Polygraph

VOLUME 29 2000 NUMBER 4

Contents

Published Quarterly
©American Polygraph Association, 2000
P.O. Box 8037, Chattanooga, Tennessee 37414-0037

Test and Retest Accuracy of a Psychophysiological Detection of Deception Test

William J. Yankee and Douglas Grimsley

Abstract

This study was designed to determine how reliable and valid a Zone Comparison PDD test is when the same individuals are administered the same test on two different occasions. Seventy-two subjects were used in an analog mock crime study using a Zone Comparison test format. Thirty-six subjects were randomly assigned to "innocent" and thirty-six to "guilty" programmed conditions. In addition, each subject was assigned to either an "accurate", "inaccurate", or "no" feedback treatment group. The overall accuracy for Tests 1 and 2 was 67% and 61% respectively. Excluding the inconclusive decisions, the accuracy for Test 1 was 94% and 88% for Test 2. Both Test results discriminated between guilty and innocent subjects. There were no significant differences: between the guilty subjects on Test 1 and Test 2; between the innocent subjects on Test I and Test 2; nor between the guilty and innocent subjects on Test 1. However, the accuracy for the guilty subjects on Test 2 was significantly different from the accuracy rate for innocent subjects on Test 2. There was no significant difference for the innocent subjects between Test 1 and Test 2, and no significant differences between feedback groups.

Key Words: Accuracy, Feedback, Reliability: Test-Retest, Validity

During the past thirty years, and particularly during the last fifteen years, a variety of validity and reliability studies have been conducted to determine the accuracy and consistency of psychophysiological detection of deception (PDD) tests[1] and their ability to discriminate between innocent and guilty subjects. These studies have involved determining accuracy rates under field conditions (using data collected in criminal investigations), laboratory conditions (using mock crimes, or numbers tests in laboratory settings absent field type procedures) and analog conditions (in laboratory settings but simulating field procedures).

Various literature reviews of PDD test studies have been published. Ansley, Horvath and Barland (1983) provided an extensive bibliography of many validity and reliability studies. More detailed information on selected studies was presented by Horvath (1976) and Raskin and Podlesny (1979). Lykken (1981), in addition to a review of picked studies, provided a critical analysis of several. Subsequent to a congressional request, a thorough analytical review of PDD processes and studies was published by the Office of Technology Assessment (1983). Ansley and Garwood (1984) presented an extensive summary and lengthy analysis of the research along with a presentation of the utility of PDD testing. In addition, the *American Psychologist* presented several articles relative to the complexities of PDD testing applications (Brooks, 1985; Katkin, 1985; Saxe, Dougherty & Cross, 1985).

Editor's Note: This paper was adapted from an earlier project. The original citation for the research is: Yankee, W., & Grimsley, D. The effect of a prior polygraph test on a subsequent polygraph test, NSA Contract NDA 904-84-C-4249, A. Madley Corporation and the University of North Carolina at Charlotte, NC. 1986. In 1998 Dr. Yankee rewrote the paper seen here for publication in *Polygraph.*

[1] "PDD tests," as used throughout this paper, are to be construed as a predetermined set of questions that are asked while collecting selected physiological data from the examinee during a broader set of examiner/examinee interactions called an "examination."

Other evaluations of selected studies have been published by Iacono and Patrick (1988), Elaad and Kleiner (1990) and Bradley and Rettinger (1992). While there are many studies addressing the general issue of the validity and reliability of PDD tests, there is a dearth of studies dealing with test and retest in which an examiner tests the same person on the same matter, on two different occasions. The majority of the studies which have been done have generally used intra-rater or inter-rater diagnostic agreement as the measure of reliability. One of the early reliability studies was reported by Rouke (1941) in which he looked at the relationship between the diagnostic opinions of an examiner on PDD examinations reviewed and analyzed at two different times (intra-rater) and the relationship of the diagnostic opinions of two independent judges (inter-rater) reviewing the same examination records. Other reliability studies by Horvath and Reid (1971), Barland (1972), Hunter and Ash (1973), Horvath (1974), and Slowick and Buckley (1975) also provide data concerning two or more independent judges evaluating the same data. Some of these studies, in addition to determining inter-rater relationships, compared the accuracy of the decisions of experienced and inexperienced examiners with the accuracy of the original examiners (Horvath, 1977; Yankee, Powell, & Newland 1988). Studies by Ellson et al (1952), Elaad (1989), Balloun & Holmes (1979), Grimsley and Yankee (1986), Yankee and Grimsley (1986) reported on the consistency of physiological responses, however, none of the studies were reasonably similar in design or testing format to warrant comparison.

A significant issue related to repeated PDD examinations on the same individual is the effect of the explicit or implicit feedback the subject receives after the first test. Does being told one is deceptive when actually truthful, or truthful when actually deceptive, or no feedback at all, after the initial examination, affect the outcome of a subsequent examination? After analyzing twelve studies that used a variety of placebo (feedback) conditions, Timm (1982) related that six of the studies he reviewed reported significant differences between the feedback condition and the effect on subsequent tests, and the other six reported no effect. However, two studies using a Comparison Question Test format evaluated the effects of feedback which was presented to the subjects before the initial test. Rovner (1986) found a decrease in accuracy on subjects that had been provided information regarding how a Comparison Question Test is conducted and then given two practice tests before the "experiment" test as compared to a group that received no information or practice; a third group received information only. Barland (1975), using field cases, found no significant differences in scores between subjects that had undergone previous examinations (involving a different issue) and those that had not.

The present study was designed to provide information about the accuracy of PDD examiner decisions after the first test as compared to the effect of various feedback conditions on the accuracy of PDD examiner decisions after the second test. The design used a test/retest procedure on the same subjects, using the same questions, from the same mock crime. An analog probable-lie comparison question specific issue PDD test was used during each examination. After the examination, one-third of the innocent and one-third of the guilty subjects were given feedback that the results of their tests were accurate; one-third of each group were given feedback that the results of their tests were inaccurate; and one-third of each group did not receive any feedback. The study provided information about: (1) overall accuracy of the examiners' decisions for Examination 1 and Examination 2; (2) a comparison of accuracy for subjects programmed guilty and those programmed innocent for each of the two tests; and, (3) the effects of accurate, inaccurate and no feedback on examiner decisions on the second test.

Method

Subjects

The subjects were 72 college students, 36 males and 36 females, recruited at the University of North Carolina at Charlotte. Each had to affirm that he or she was at least 18 years of age or older and had never had a PDD examination before, either real or experimental. Each was given credit toward course requirements and paid $15 for participation.

Apparatus

The tests were conducted in an electrically-shielded room which is part of the psychophysiological laboratory complex maintained by the Department of Psychology. The room was equipped to maintain normal temperature and humidity levels.

An Executive UltraScribe Stoelting Polygraph was used to record the selected physiological activities. Five recording channels were used: two electronically enhanced pneumograph units; two electrodermal response (EDR) recorders; and one electronically enhanced cardiovascular recorder. One pneumograph tube was placed around the upper chest and the other around the abdomen. One EDR channel used regular field electrodes, and the other used silver/silver chloride electrodes with a potassium chloride and agar mixture as the conducting medium. The EDR field electrodes were placed on the index and ring fingers of one hand and the EDR silver/silver chloride electrodes were placed on the index and ring fingers of the other hand. These were reversed on each odd numbered subject. The blood pressure cuff was placed on the left arm and inflated sufficiently to produce a satisfactory record of cardiovascular activity at low sensitivity levels. Customary attachment of the sensors to the subject, and normal instrument calibration procedures were used during all phases of data collection.

Examiner Qualifications

The examiner received his PDD examination training at an institute accredited by the American Polygraph Association. He possessed a Bachelor of Arts degree and had completed a substantial amount of work on a Masters degree. He had three years of experience as a police officer and eight years as a PDD examiner. He taught as a regular instructor at an accredited polygraph school for four years and presented special instructional programs at polygraph colloquia, and various state and federal PDD association meetings.

Procedures

The 72 subjects were randomly assigned to the experimental conditions. Each subject was involved in the experiment on three consecutive days. On the first day, as a group, they were given a general orientation to the study. They were told that some of them would be asked to commit a mock crime and then lie about it during a PDD examination, while others would take the same PDD examination but would be telling the truth when they denied being involved in a mock crime. They were advised that they would all be taking two PDD examinations on two consecutive days. Subsequent to the group orientation, individual instruction sessions were given. During these individual sessions the subjects were advised regarding the condition to which they were assigned, were provided with "Information for Human Research Participants" and after reading this and agreeing to continue, signed an "Informed Consent Form". They were also told that they would receive course credit and $15 for participating.

The subjects assigned to the "innocent" condition were told about the mock crime in general terms as follows: someone went into a faculty member's office, unlocked a desk drawer, and took a digital watch. They were advised to be completely truthful about not participating in the mock crime when they were questioned by the examiner before and during their PDD examinations.

Subjects assigned to the "guilty" condition were put through a mock crime scenario. They were told to go to a designated office, to enter the office, close the door, find a ring of keys on the desk top, find the key that fit the lock on the right top desk drawer, unlock the drawer, observe a gold-colored digital watch, take the watch from the drawer and put it in their pocket or someplace on their person where it could not be seen, relock the drawer, put the keys back in their original location, and leave the office, shutting the door behind them. They were told to avoid being seen going in and out of the office, but if they were challenged by anyone, to say they made a mistake and continue on their way. From a clandestine observation post, an accomplice of the experimenters approached each subject as they left the office and curtly asked, "What were you doing in my office?" but made no attempt to question the explanation given by the subjects.

All subjects met with the controller before their scheduled test each day. The innocent subjects were reminded to be truthful in their answers to all the questions they would be asked by the examiner. The guilty subjects were also told to be truthful in answering all questions asked by the examiner except questions about the office, the keys, the drawer, or the watch. To these questions they were told to lie and not reveal in any way that they had been involved in the mock crime.

The pretest interview approximated normal field procedures, and involved explaining what was going to take place, how the examiner would proceed, how the autonomic nervous system functions when lying, how the instrument recorded the physiological activities, a review of all the questions to be asked, including the field approach to establishing comparison questions. The Zone Comparison Technique format was used. The questions were asked once during each of the three test runs, during each of the two examinations.

The tests consisted of the following questions:

1. Is your last name ___________?

2. Regarding that missing digital watch, do you intend to answer truthfully to each question about that?

3. Are you completely convinced that I will not ask you a question during this test that has not already been reviewed?

4. Prior to ______, did you ever take anything that did not belong to you?

5. Did you take that missing digital watch?

6. Prior to _____________did you ever take someone's property without permission?

7. Did you take that digital watch from that locked desk in the manager's office?

8. Is there something else you are afraid I will ask you about even though I told you I would not?

9. Prior to __________did you ever tell a lie to stay out of trouble?

10. Do you know for certain who took that missing digital watch?

After the first examination (running the test three times), the examiner told the subject he was going to leave the room to analyze and diagnose the collected data. The examiner did not, however, diagnose the data at that time. Rather, he went to a pre-arranged location where the controller had left a slip of paper containing the appropriate statement to make to the subject. This statement was based on randomly assigned treatment for each subject of accurate feedback, inaccurate feedback, or no feedback. The examiner told the subject whatever the diagnosis was as written on the slip. The subjects were then advised to return to the controller.

On the next day the subjects were briefed again regarding their appropriate roles, and then proceeded to the examination room to take their second test. The second test was conducted the same as the first test in all respects except that a stimulation test was conducted before the first CQT test was administered. After the second examination (running of the test three times) the subjects were instructed by the examiner to return to their controller and that he would call the controller to report his diagnosis.

Evaluation of Physiological Data

The physiological analysis criteria (not the scoring procedures) used by the Department of Defense Polygraph Institute was applied by the examiner. A 7-position scale of numbers from +3 to -3 was used to represent the perceived response differences between the relevant and comparison questions for each zone (Questions 5, 7, and 10 were relevant questions and 4, 6, and 9 were comparison questions.) A negative number indicated that the reactions to the relevant question were greater than those to the comparison question, and a positive number indicated the opposite. A "0" was assigned when the examiner could not perceive a difference in the reactions to the comparison and relevant questions.

Results

The overall accuracy of the examiner decisions for Test 1 and Test 2 are shown in Table 1. The examiner's decisions resulted in

48 correct, 3 incorrect and 21 inconclusive for Test 1, and 44 correct, 6 incorrect, and 22 inconclusive for Test 2. Thus, the overall accuracy rate was 67% for Test 1 and 61% for Test 2. If the inconclusive results are removed, since they are actually a suspension of judgment, the accuracy achieved for the decisions made was 94% for Test 1 and 88% for Test 2. A 3X2 chi square test analysis showed no significant difference between the distributions (X^2=1.1971, df=2, p=.5495). Since the number incorrect decisions was small (3), the incorrect and inconclusives cells were collapsed and a Yates correction applied again resulted in no significant difference (X^2=.2709, df=1, p=.6027).

Table 1

Accuracy of Examiner Decisions on Tests 1 and 2

	Test 1 (n=72)		Test 2 (n=72)	
Diagnosis	n	%	n	%
Correct	48	67	44	61
Incorrect	3	4	6	8
Inconclusive	21	29	22	31
Correct Excluding Inconclusives		94		88

The overall accuracy of the examiner's decisions for guilty and innocent subjects from the two examinations were explored. Table 2 shows the overall accuracy data for the guilty subjects for Test 1 and Test 2. It can be seen that the examiner was correct on 20 decisions and incorrect on 3, while 13 of the outcomes were inconclusive for Test 1. The number of accurate decisions dropped to 16, the number of incorrect decisions increased to 5, while inconclusives increased to 15 on Test 2. Removing the inconclusive finding resulted in accuracy figures of 87% and 76% for Test I and Test 2, respectively. There was no significant difference among these data (X^2=1.087, df=2, p=.5806). Because of the small values in the two "incorrect" cells, the "incorrect" and "inconclusive" cells were collapsed and a Yates correction applied. Again, there was no significant difference (X^2=5, df=1, p=.4795).

Table 2

Accuracy for Guilty Subjects on Tests 1 and 2

	Test 1 (n=36)		Test 1 (n=36)	
Diagnosis	n	%	n	%
Correct	20	56	16	44
Incorrect	3	8	5	14
Inconclusive	13	36	15	42
Correct Excluding Inconclusives		87		76

Table 3 shows comparable data for the innocent subjects for Test 1 and Test 2. The examiner was exceptionally accurate in making these decisions. There were 28 correct and 0 incorrect decisions, while 8 outcomes were inconclusive on the first test. Test 2 data resulted in 28 correct and 1 incorrect decisions, with 7 inconclusive results. If the inconclusive results are removed, the examiner was correct 100% of the time for the first test and 96% for the second test with the subjects programmed innocent. The analysis of these data using a 3X2 chi square failed to show any significant differences (X^2=1.066, df=1, p=.7768). The accuracy rate for guilty compared to innocent for Test 1 and the accuracy rate for guilty compared to innocent for Test 2 were explored.

Table 3

Accuracy for Innocent Subjects on Tests 1 and 2

	Test 1 (n=36)		Test 1 (n=36)	
Diagnosis	n	%	n	%
Correct	28	78	28	78
Incorrect	0	0	1	3
Inconclusive	8	22	7	19
Correct Excluding Inconclusives		100		96

Table 4 provides information regarding the guilty and innocent for Test 1. It can be observed that the examiner was correct on 20 and incorrect on 3 of the decisions and had 13 inconclusive outcomes on the first examination for the guilty subjects. With the innocent subjects, the examiner was correct in 28 decisions, had no incorrect decisions but had 8 inconclusive results. An analysis of these data using a 3X2 chi square failed to show any significant difference (X^2=5.5238, df=2, p=.06317). Collapsing the incorrect and inconclusive cells and applying the Yates correction in a 2X2 chi square also failed to show a significant difference (X^2=3.0625, df=1, p=.0801).

Table 4

Accuracy for Guilty and Innocent Subjects on Examination 1

	Guilty (n=36)		Innocent (n=36)	
Diagnosis	n	%	N	%
Correct	20	56	28	78
Incorrect	3	8	0	0
Inconclusive	13	36	8	22
Correct Excluding Inconclusives		87		100

Table 5 provides data as it relates to accuracy rates for guilty and innocent subjects for Test 2. It can be observed that the examiner made 16 correct and 5 incorrect decisions with 15 inconclusive results with the guilty subjects while making 28 correct and 1 incorrect decision with 7 inconclusive results with the innocent subjects. A 3X2 chi square resulted in determining a significant difference between the two distributions (X^2=8.8484, df=2, p=.01). However, a 2X2 chi square (collapsing the incorrect and inconclusive cells) using a Yates correction resulted in no significant difference (X^2=7.0714, df=1, p=.0783) in the data.

Table 5

Accuracy for Guilty and Innocent Subjects on Test 2

	Guilty (n=36)		Innocent (n=36)	
Diagnosis	n	%	n	%
Correct	16	44	28	78
Incorrect	5	14	1	3
Inconclusive	15	42	7	19
Correct Excluding Inconclusive	76		97	

The issue of feedback, provided to the subjects prior to the second test, was also examined. The accuracy rates for each group on each test can be observed in Table 6. Using a 3X2 chi square with Yates correction resulted in no significant difference between the distribution of accurate, inaccurate and no feedback groups for Test 1 (X^2=3.375, df=2, p=.2849) and for Test 2 (X^2=.8181, df=2, p=.6643). A 2X2 chi square analysis with Yates correction comparing Test 1 and Test 2 for each of the categories resulted in no significant differences for the accurate group (X^2=.9076, df=1, p=.3407), inaccurate group (X^2=.0937, df=1, p=.7594) and no feedback group (X^2=.7720, df=1, p=.7721).

Table 6

Accuracy for Accurate, Inaccurate and No Feedback Groups for Examinations 1 and 2

	Accurate				Inaccurate				No Feedback			
	Test 1		Test 2		Test 1		Test 2		Test 1		Test 2	
Diagnosis	n	%	n	%	n	%	n	%	n	%	n	%
Correct	19	79	15	63	16	67	16	67	13	54	13	54
Incorrect	0		2	8	2	8	2	8	1	4	2	8
Inconclusive	5	20	7	29	6	25	6	25	10	42	9	38
Correct Excluding Inconclusive	100		88		89		89		93		87	

Discussion

The overall accuracy rate for the examiner (excluding inconclusive decisions) is very high and consistent with the general body of PDD literature for studies involving the comparison question test format. The fact that on Test 1 there were 48 correct and 3 incorrect decisions, and on Test 2 there were 44 correct and 6 incorrect decisions, provides evidence that supports the efficacy of PDD testing procedures in detecting deception when examinations are repeated. Considering that this was an analog study, there were a reasonable number of inconclusive findings.

The issue of whether innocent or guilty subjects are more easily detected by the PDD procedures appears to be a controversial one. Generally, the literature shows that guilty subjects are more readily detected than innocent subjects in both field and analog studies (Barland, 1972; Hunter & Ash, 1973; Horvath, 1974; Barland, 1972; Slowick & Buckley, 1975; Wicklander & Hunter, 1975; Horvath, 1976; Lykken, 1981; Stern, Breen, Watanabe & Perry, 1981; Patrick & Iacono, 1989, 1991). The results are not unanimous. For example, truthful subjects were more frequently identified in the study by Balloun and Holmes (1979); in one phase of a study by Waid, Orne and Orne (1981); but equal in rates of detection in Podlesny and Raskin (1978); Raskin and Hare (1978); Rovner, Raskin and Kircher (1979). However, the bulk of the evidence seems to support that guilty subjects are more readily identified. The vast differences between lab and field studies, the variety of experimental designs and the manifold methodologies used, make comparisons of the results difficult. The high accuracy obtained on the innocent subjects in this study may have been the result of the relevant and personal nature of the probable-lie comparison questions overriding the relatively innocuous relevant questions associated with the mock crime. For example, the comparison question, "Prior to ____, did you take anything that didn't belong to you?" relates to real life matters and could be very significant; whereas, the relevant questions associated with a mock crime in a role-playing scenario, "Did you take the missing digital watch?" may be of little consequence for some subjects in relation to the comparison question. Thus, in scoring, the numerical values could be weighted in favor of the innocent subject in a mock crime study. In real life testing, the effect may be different, for both the guilty and the innocent.

The absence of any significant effects attributable to the feedback (accurate, inaccurate or no feedback) is somewhat surprising. The number of correct decisions declined for all groups on Test 2 but was more pronounced for the no feedback group. The data suggest that giving inaccurate feedback to subjects does not reduce correct decisions but, since the numbers were so small in each of these categories, additional work needs to be done concerning the effects of feedback on repeated PDD examinations. Since subjects are frequently given repeated examinations under field conditions (with implied or explicit accurate, inaccurate or no feedback), additional studies, analog and field, need to be carried out to verify and extend the present findings.

References

Ansley, N., & Garwood, M. (1984). The accuracy and utility of polygraph testing. *Polygraph,* 13, 3-131.

Ansley, N., Horvath, F., & Barland, G. (1983). *Truth and Science: A Bibliography, 2d Ed.,* Maryland American Polygraph Association.

Balloun, K.D. & Holmes, D.D. (1979). Effects of repeated examinations on the ability to detect guilt with a polygraphic examination: A laboratory experiment with a real crime. *Journal of Applied Psychology,* 64, 316-322.

Barland, G. (1972, August). The reliability of polygraph chart evaluation. Presented at the American Polygraph Association Annual Meeting, Chicago, IL.

Barland, G. & Raskin, D. (1975). An evaluation of field techniques in detection of deception. *Psychophysiology,* 12(3), 321-330.

Bradley, M.T. & Rettinger, J. (1992). Awareness of crime-relevant information and the Guilty Knowledge Test. *Journal of Applied Psychology,* 77(1), 55-59.

Brooks, J. (1985). Polygraph testing: Thoughts of a skeptical legislator. *American Psychologist,* 40, 348-354.

Edel, E.D. & Jacoby, J. (1975). Examiner reliability in polygraph chart analysis: Identification of physiological responses. *Journal of Applied Psychology,* 60, 632-634.

Elaad, E., & Kleiner, M. (1990). Effects of polygraph chart interpretation experience in psychophysiological detection of deception. *Journal of Police Science and Administration,* 17(2), 115-123.

Horvath, F.S. (1974). The Accuracy and Reliability of Police Polygraphic (Lie Detector) Examiner Judgments of Truth and Deception: The Effect of Selected Variables, Unpublished Doctoral Dissertation, Michigan State University.

Horvath, F.S. (1976). Detection of deception: A review of field and laboratory research. *Polygraph,* 5, 107-145.

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

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

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

Iacono, W.G., & Patrick, C. (1988). Assessing deception: Polygraph techniques. Reprinted from *Clinical Assessment of Malingering and Deception,* Edited by Richard Rogers, Guilford Press: New York.

Katkin, E. (1985). Polygraph testing, psychological research, and public policy: An introductory note. *American Psychologist,* Vol. 40, 146-47.

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

Lykken, D.T. (1981). *A Tremor in the Blood: Uses and Abuses of the Lie Detector,* New York: McGraw-Hill.

Patrick, C., & Iacono, W.G. (1989). Psychopathy, threat and polygraph test accuracy. *Journal of Applied Psychology,* 74(2), 347-355.

Patrick, C., & Iacono, W.G. (1991). A comparison of field and laboratory polygraph in the detection of deception. *Psychophysiology,* 28(6), 632-638.

Podlesny, J.A., & Raskin, D.C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology,* 15, 344-359.

Raskin, D.C. (1986). The polygraph in 1986: Scientific professional and legal issues surrounding application and acceptance of polygraph evidence. *Utah Law Review,* 29(1), 29-74.

Raskin, D.C., & Podlesny, J.A. (1979). Truth and deception: A reply to Lykken. *Psychophysiological Bulletin,* 86, 54-58.

Rouke, F. (1941). Evaluation of the indices of deception in the Psychogalvanic Technique, Unpublished doctoral dissertation, Fordham University.

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

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

Saxe, L., Dougherty, D., & Cross, T. (1985). The validity of polygraph testing. *American Psychologist*, 40, 355-366.

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

Stern, R.M., Breen, J.P., Watanabe, T., & Perry, B.S. (1981). Effect of feedback of physiological information on responses to innocent associations and guilty knowledge. *Journal of Applied Psychology*, 66, 677-690.

Timm, H.W. (1982). Analysis deception from respiration patterns. *Journal of Police Science and Administration*, 10, 47-51.

U.S. Congress, Office of Technology Assessment (1983). Scientific Validity of Polygraph Testing (OTA-TM-H-15), Washington, D.C., Government Printing Office.

Venables, P., & Christie, M. (1973). Mechanisms, instrumentation, recording technique and quantification of responses. In W.F. Prokasy and D.C. Raskin (Eds.), *Electrodermal Activity in Psychological Research*, New York: Academic Press.

Waid, W., Orne, E., & Orne, M. (1981). Selective memory for social information, alertness and physiological arousal in the detection of deception. *Journal of Applied Psychology*, 66, 224-232.

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

Yankee, W., Powell, J., & Newland, R. (1988). An investigation of the accuracy and consistency of polygraph chart interpretation by inexperienced and experienced examiners. *Polygraph*, 14, 107-117.

Yankee, W., & Grimsley, D. (1986). The Effect of a Prior Polygraph Test on a Subsequent Polygraph Test, NSA Contract NDA 904-84-C-4249, A. Madley Corporation and the University of North Carolina at Charlotte, NC.

He Said / She Said: Polygraph Evidence In Court

Jonathan Marin

Key words: commentary, evidentiary polygraph

In recent years, frequent instances of wrongful criminal convictions and mistaken verdicts in prominent civil cases have attracted media attention, and become a matter of public concern. They have reinforced the sense that attorneys, especially prosecuting attorneys, often willingly and even eagerly make use of highly suspect testimony. The Ramparts police perjury scandals in Los Angeles, and similar scandals in Illinois, West Virginia, and other states, have heightened public awareness of the untrustworthiness of testimony by police and informants, especially jailhouse informants and accomplices testifying under plea agreement.

This article looks at how the polygraph might be used to address the problem. It considers issues of validity and reliability, and rejects the traditional positions of both proponents and opponents of the polygraph. Instead, it presents the case for an innovative, but sound, "third way." Some constituencies of both groups may perceive the proposal as a threat to their interests. Nevertheless, I think it is a reform worth fighting for, in order to restore public confidence in testimony by police officers and cooperating witnesses and to prevent miscarriages of justice.

The article addresses the issues and problems--scientific, legal, and social--that surround the use of polygraph evidence in court and suggests an approach to its use that I think safely navigates the minefield they present. It presents the idea that polygraph results should not be admitted as evidence in their own right but rather should be used as a tool to screen out untrustworthy evidence.

Courts should exclude testimony from a witness who has tested "deceptive" where--and only where--that result is corroborated by a "nondeceptive" result on the opposing side ("paired results"). Using only paired results resolves the unquantifiable uncertainties and reduces by a factor of at least 5 the quantifiable uncertainties that underlie most legitimate resistance to the polygraph. If individual examination results are incorrect 20% of the time, the chance that two together will be in error is only 4%. Juries make mistakes. Evidence that has at most a 4% likelihood of being true is too untrustworthy to warrant submitting it for their consideration. Excluding it will reduce the incidence of perjury and the number of mistaken verdicts. It will discourage frivolous claims and frivolous defenses, thereby reducing court caseloads and backlogs. Wherever possible, the decision to exclude should be made before trial.

The article addresses broad problems of admissibility and exclusion of evidence affecting both civil and criminal cases. In respect to criminal cases, it considers the constitutional implications of the approach, particularly those having to do with the right against self-incrimination. It also analyzes alternative approaches and their drawbacks.

Introduction

In many court cases, civil as well as criminal, the two sides present witnesses whose factual claims clash. Both cannot be telling the truth. A woman swears she was raped; the defendant swears it was consensual. An arrested suspect charges that

Reprinted with permission. The original can be found at http://users.rcn.com/jonmarin/Polygraph1.htm. Reprint requests should be directed to Jonathan Marin, P.O. Box 840, Brooklyn, NY 11202.

police used excessive force; the officers deny it. A jailhouse informant swears that the defendant confessed a murder to him; the defendant swears it isn't true. In these situations, polygraph results may prove invaluable, if used correctly.

Polygraph evidence is no longer barred from the courtroom. The Supreme Court has left it to the courts of each jurisdiction to determine how and when to allow it, or to exclude it altogether [*United States v. Scheffer*, 523 U.S. 303 (1998)]. I believe that the courts' safest, simplest, and most productive use of the polygraph is to exclude testimony about a fact from any witness who has tested "deceptive" about that fact whenever a witness from the opposing side has tested "non-deceptive" about that fact. No jury would ever hear, or hear of, the polygraph results themselves. A witness's refusal to submit to a polygraph examination on any factual claim would be treated as a "deceptive" result in regard to that claim in civil cases; it would be treated similarly in criminal cases except where the refusing witness is the defendant.

A defendant would have the right to demand that a jailhouse informant be polygraphed concerning an alleged confession, and to be polygraphed himself. If the informant refused to take the test, his testimony would be inadmissible. If the informant tested positive for deception, and the defendant negative, then the informant's testimony would be inadmissible. If the defendant did not demand that the informant be tested, or the test produced some other combination of results, the informant's testimony would be admissible, and, except that neither side would be allowed to make any reference to the polygraph, he would be subject to cross-examination as any other witness.

Polygraph results are unreliable to some degree. How a polygraph chart is interpreted can depend on the thresholds of physiological variance above which a response will be called "deceptive" and below which it will be called "nondeceptive." (Responses falling between the thresholds are called "inconclusive.") It will be the court's responsibility to determine that the thresholds, the competence of examiners, and the conditions under which tests have been given are in accordance with the standards enforced by federal agencies--or at least the recommendations of the American Polygraph Association. Results should be accompanied by an unedited beginning-to-end videotape of the examination.

Reliability, Validity, and the Power of Pairing

Recent studies that have sought to quantify polygraph tests' accuracy have estimated both false positives and false negatives at just under 10% [Honts, *US v Scheffer* (Amicus)]. "Accuracy" can be misleading, however, especially where the proportion of subjects who are in fact telling the truth is high. Suppose that from a sample of 100 subjects, of whom only 1 is "deceptive," a test found 2 of the subjects to be "deceptive." It could claim 99% accuracy (if one of them was the deceptive subject). Impressive, but the likelihood that a failing subject had actually been deceptive would be only 50%. The field studies cited in Honts (average false positive rate = 9.5%) imply that the 50% figure probably does approximate the ratio of false positives to true positives in the real world of criminal investigations.

Drawing correct inferences from stand-alone results requires knowledge about the relevant samples that is rarely possible outside the laboratory and an understanding of statistical inference that is beyond the experience of jurors. Admitting single results can easily clear the guilty and imperil the innocent. When two people dispute a fact within the personal knowledge of both, however, usually one is telling the truth and the other is lying. Pairing results therefore assures the balanced samples necessary to support sensible inferences. Even allowing for a modest percentage of witnesses who are honestly mistaken, and of cases where both are lying, the known accuracy of the test can be safely applied. When results are paired and the second result confirms the first, then according to probability theory, the likelihood of an erroneous conclusion is the product of the two individual probabilities. Supposing the tests' individual probability of error to be as high as 20%, the probability that confirmed "deceptive" testimony would be true would be only 4% (0.2 x 0.2 = 0.04 = 4%), and the probability that it would be false would therefore be 96% (100 - 4 = 96%).

Courts continue to agonize over whether to accept polygraph results as "scientific" and admit them into evidence. But information doesn't have to go before a jury (or other finder of fact) in order to be useful. Using the polygraph to exclude testimony that has a 96% likelihood of being false accords both with common sense and with the Supreme Court's view that "Exclusion... is usually premised on the view that admission would lead to the frequent presentation of perjured testimony to the jury" and that "untrustworthy evidence should not be presented to the triers of fact" [*Chambers v Mississippi*, 410 U.S. 284 (1973)]. Excluding such testimony would not usurp the role of the jury as ultimate fact-finder any more than such time-honored exclusions as the rule against hearsay, and doing so could well prove as valuable as the hearsay rule in steering juries away from mistaken results.

There are at least two quite distinct purposes that polygraph evidence can serve in court. One is to present negative ("nondeceptive") test results in order to bolster the credibility of witnesses. The other is to present positive ("deceptive") results in order to preclude witnesses from testifying or impeach their credibility. Both arise from the same technology, but the scientific and statistical bases for trusting them, and the practical and legal considerations surrounding them, differ greatly.

Through the years, the primary focus of the polygraph debate has been the admissibility of individual "nondeceptive" polygraph results to bolster testimony, especially that of criminal defendants and prisoners whose test results point to their innocence. The points raised by both sides focus on the trustworthiness of polygraph results treated on a stand-alone basis. The proponents of wider use have argued for admitting test results as trial evidence. Admissibility is a difficult argument to win, and its proponents have rarely been successful. Results could be admitted only after an elaborate, tedious, and time-consuming courtroom minuet. They would have to be supported by the examiner, and perhaps other experts, as well as be subjected to challenge by cross-examination and the presentation of contrary evidence, and to a web of instruction, some of it highly technical, by the court. Exclusion based on paired tests circumvents those difficulties. Because of the benefits it offers to police, prosecutors, courts, defense and civil bar, and honest parties, it is the approach that provides proponents of widening the courts' use of the polygraph their best prospect of success.

Acceptance of paired results can help free many wrongly convicted prisoners, whereas stand-alone results face an insurmountable public acceptance problem. Suppose that 90% of prisoners are guilty of the crimes for which they are incarcerated and that false negatives average about 10%. Then for every 100 prisoners tested, there would be 80 true positives, 10 true negatives, 9 false negatives, and 1 false positive. About half the people that stand-alone tests would release would in fact be guilty. Acting where prisoners test negative and their accusers test positive would reduce that to 10%. Setting free 10 innocent persons, at the price of freeing 1 guilty one, is a supportable, achievable objective.

Stand-Alone Negative Results

The case--scientific, legal, and social--against allowing negative ("nondeceptive") results on a stand-alone basis is strong. The statistical underpinning of negative results is problematical because of the difficulty of quantifying the false negatives in the absence of "ground truth"--an external yardstick by which to measure whether subjects are deceptive. In field work, ground truth is notoriously difficult to determine.

There is no straightforward way to ascertain false negative rates--the percentage of subjects testing "nondeceptive" who were in fact deceptive--in real-world samples. To be a known false negative, a subject must first beat the test and later be found out. That rarely happens. When people beat the test, it usually remains their secret. It is not known how many cases go unsolved because a false negative was excluded from further investigation and how many because the culprit was not among those tested. In laboratory tests, false negative rates are usually about 10%, but extrapolating them to the real world is difficult. The physiological changes the equipment measures are affected by the subjects' fear. The higher the stakes, the greater the fear of being caught in a lie, and the greater the measured response. However expert the examiner and well conducted the test, the high stakes of real-world tests cannot be duplicated in the laboratory.

Allowing stand-alone "nondeceptive" polygraph evidence is fraught with other difficulties. Once it were allowed, litigants would seek to introduce polygraph evidence to buttress many, even most, witnesses. Juries would come to expect them to do so. Since polygraph results cannot be introduced into evidence without the testimony of the examiner or other expert to interpret them, this would mean a de facto return to the archaic voucher system of the Middle Ages, when litigants were expected to produce "voucher witnesses" to vouch for the credibility of their witnesses.

The parade of voucher witnesses would tie up dockets and, by lengthening trials, would add to the cost of litigation for all parties. Moreover, polygraph examinations are expensive, and examiners are well paid for their time in court. In civil cases, the "voucher effect" would tend to raise the price of justice, aggravating the already serious disadvantage faced by parties with limited budgets. It would be especially pernicious in criminal trials, where strategy considerations often preclude defendants' taking the stand. Allowing "nondeceptive" results into evidence in support of prosecution witnesses would practically compel criminal defendants to be polygraphed and testify, giving prosecutors an unacceptable subterfuge around the right against self-incrimination.

Many possible countermeasures that would enable deceptive subjects to fool polygraph examiners have been suggested. Their utility remains unproven, but to the extent they may be or become effective, their use would affect only negative results. Police officers testify frequently, and they are trained to do it effectively. Many are professional witnesses. If effective countermeasures could be mastered, unscrupulous police and other "professional witnesses" would be among the first to learn them. The danger exceeds their numbers of such witnesses because of the number of times each would testify through the course of a career.

Special protections afforded criminal defendants introduce a selectivity bias into the process. If defendants had the option of introducing polygraph evidence, their counsel could be counted on to bury unfavorable results. Defendants would take tests privately-a no-risk option. Most defendants opting for private tests would presumably fail. The court would never know. Only defendants with favorable test results would introduce them. If defendants in, say, 10% of trials presented "nondeceptive" test results, would that mean that 10% of defendants are innocent? That the polygraph is subject to 10% false negatives? This selectivity bias applies primarily to criminal defendants, not to most other witnesses. But no court that allowed stand-alone negative results to bolster the testimony of some witnesses could constitutionally bar criminal defendants as a class from using them. To the extent false negatives occur, the selectivity bias would lead to wrongful acquittals.

James K. Murphy, the former polygraph unit chief at the FBI laboratory in Washington, D.C., has testified (http://truth.boisestate.edu/polygraph/MURPHY1.HTML) that the FBI annually administers polygraph examinations to about 5,000 applicants for sensitive jobs. Each applicant takes two tests. Applicants almost always pass the first test, which focuses on counter-intelligence issues: Applicants are asked whether they've ever been in contact with anybody from a foreign intelligence service and whether they were directed to seek FBI employment. The failure rate is about 0.5%.

The applicants' charts from the first test are used for comparison with the charts from their second test, which deals with use of illegal drugs, abuse of legal drugs, and falsification of the application for employment. In accord with ordinary knowledge and common sense, the failure rate on the second test is much higher: More applicants have had undesirable experience with drugs than have an involvement with espionage. More than 70% of applicants failing the second test have validated the examination results through confession or through admission at the time of the test.

The FBI believes that these results support validation, through the correspondence of the results with the known statistical base rates for those two subject areas, and achieve reliability as the test relates to them. They rely heavily on these results, notwithstanding that the test results provide only a weak inference regarding false negatives.

Despite the scientific and statistical difficulties with "nondeceptive" results, federal, state, and local police and prosecutors place great confidence in them and make important decisions based upon them. The methodological argument against their use on a stand-alone basis is not that they are valueless, but that their value is so uncertain. The rub is that the testimony of interested parties, informants, and plea-bargained accomplices is also uncertain.

Stand-Alone Positive Results

The methodological, scientific, and statistical grounds for confidence in estimates of the rate of false positives are stronger. The FBI, OSI, and CIA have administered polygraph examinations to tens of thousands of past, present, and prospective government employees and armed forces personnel. The 0.5% failure rate cited by Mr. Murphy of the FBI indicates that when tests are given under proper conditions by competent examiners, and interpreted using a high threshold of physiological variance, false positives, taken as a percentage of tests administered, can be extremely low. This low occurrence of positive results occurs in a real-world setting where the stakes for the examinees are not only their jobs, but also the unpleasantness of becoming the subject an espionage investigation. Since there obviously cannot be more false positives than there are positives, the percentage of positive results establishes a rigorous limit for those thresholds. But for purposes of evaluating the significance of an individual positive result, it is the ratio of false positives to all positives that matters. If there is one spy among a population of subjects, and two (including the culprit) fail the test, then that ratio is 50%, irrespective of the number of subjects in the sample.

In testing conducted by police for the purpose of eliminating possible suspects, where the subjects are people who have an appreciable likelihood of being involved in a crime, positives occur more often. Many subjects who get positive results confess and provide independent evidence that supports their confession or are convicted by juries with no knowledge of the polygraph results, thereby reducing the number of positives that might be false positives and helping scientists further refine their estimates of the trustworthiness of positive test results.

Nevertheless, I believe that no testimony from a witness who has tested "deceptive" should be excluded unless a contradicting witness has tested "non-deceptive." The second result increases confidence, perhaps by a factor of 5 or more, that the excluded testimony is really untrustworthy. Both witnesses may be untruthful, and no advantage should accrue to the one who has refused to be tested. In criminal cases, defendants would be freed from the Hobson's choice of having to testify before the jury in order to contradict testimony they know to be false.

Police And Prosecution Issues

Police and prosecutors have consistently opposed allowing polygraph results into court. The polygraph is an extremely useful investigative tool that enables them to screen possible suspects and focus their resources effectively. Police cannot compel suspects to take polygraph tests, due to the rule against self-incrimination. If failing results could be introduced into court, even many innocent people would be reluctant to risk consent, and the police would lose an invaluable time-saver. If, in order to keep the tool, they were to promise not to use results in court, their relation to the technology would remain exactly as it is now. Only defendants would stand to benefit from admissibility. Even if acknowledging a refusal to take the police polygraph became a condition of defendants' introducing "nondeceptive" results, such refusals would be credibly explainable, in light of the favorable test result, as due to an innocent person's distrust of the police.

By limiting prosecutors' use of a positive result to exclusion, the subjects' risk is reduced and the legitimate police concern addressed. Their stated concerns no doubt mask the unwillingness of some law enforcement people to forgo the advantage they gain from police perjury and other dubious testimony. To the extent that the polygraph removes that advantage, it removes a blight. The Ramparts scandals in Los Angeles, and similar scandals elsewhere, come into an atmosphere of increasing public awareness of wrongful convictions. Together, they threaten to foster a deep and long-lasting suspicion against testimony by police and informants, especially jailhouse informants and accomplices testifying under plea agreement.

It is no mere public relations problem. It is a serious cloud, and it will require concrete measures to dispel it. Apart from problems arising from a generalized negative attitude toward police, there is the specific danger of lost convictions due to excessive juror skepticism. I think the use of polygraph results suggested here is a reform that will help to restore public confidence in testimony by police officers and cooperating witnesses. Departments that opt for the idea will find allies among the media and among groups that are opposed to prosecutorial misconduct and wrongful convictions, whatever their attitude toward police image problems.

Conclusion

Imagine that a couple of King Arthur's knights have arrived here in a time machine, and happened upon a refrigerator. One of them suggests using it as a boat, while the other says no, let's fill it with dirt and grow vegetables. To me, the debate over the use polygraph results in court feels much the same. One side wants to bar them altogether, while the other wants the results of single stand-alone tests to be admitted before juries. They are equally incorrect. It is a third course--using paired results to screen testimony--that I believe is the right one.

Paired polygraph results can streamline court proceedings, lighten prosecution and legal assistance caseloads, reduce litigation costs, and move the docket along. Countless laborious cross-examinations will never take place. Many witnesses will not appear at all. Many fraudulent and frivolous cases will never be brought, and many others will never reach trial.

Ensuring that the utilization is proper is nontrivial but straightforward. Judges need to know about proper polygraph examination procedure and interpretation of results. They need to be able to reject unqualified and "bought" examiners, and conclusions based on unreasonable thresholds. Judges are capable of learning this and applying the knowledge on an ongoing basis. Unlike juries, whose secret deliberations give few clues as to the weight given to admitted evidence, judges' decisions are open, would apply to specific testimony, and would be based on examination charts and videotapes that will become part of the record, making their decisions subject to review.

The approach offers the prospect of reducing the number of wrongful criminal convictions, and of wrongful outcomes of civil cases, in a way that skirts a potential minefield of technical uncertainties as well as legal and social complications. The courts of every jurisdiction would do well to consider it.

Efficacy of Repeated Psychophysiological Detection of Deception Testing

Andrew B. Dollins, Victor L. Cestaro, and Donald J. Pettit

Abstract

Physiological measures were recorded during repeated psychophysiological detection of deception (PDD) tests to determine if physiologic response levels change with test repetition. Two groups of 22 healthy male subjects completed six Peak of Tension PDD tests on each of two test days. A minimum between test day interval of six days was maintained. The treatment group was programmed to respond deceptively to one of seven test questions while the control group was programmed to respond truthfully to all questions. The respiration and Galvanic Skin Response (GSR) line lengths, GSR peak response amplitude and latency, and cardiovascular inter-beat-interval (IBI) were calculated for each response. Analyses indicated that: except for GSR peak response latency, differential physiological reactivity during a PDD test did not change significantly during repeated tests or days; there was a decrease in average respiration line lengths during the beginning test(s) of each day; and, differential changes in average respiration line length, GSR peak latency, and cardiovascular IBI responses corresponded to deception. Power analyses are presented to assist in result interpretation. It is suggested that PDD decision accuracy, concerning subject veracity, should not decrease during repeated testing. It is further suggested that pneumograph line length and cardiovascular IBI are reliable response measures which may be sensitive to physiological changes associated with deception.

Key Words: electrodermal response, galvanic skin response (GSR), heart rate, peak of tension (POT), power analysis, psychophysiological detection of deception (PDD), repeated measures, respiration

The United States Department of Defense, various law enforcement agencies, and officers of the court routinely use a psychophysiological detection of deception (PDD) examination to determine an individual's truthfulness concerning topics of interest (Office of Technology Assessment, 1983, pp. 1-8; Lykken, 1981, pp. 1-4). The theory underlying PDD is that physiologic reactivity, in response to the presentation of a stimulus, varies with the personal relevance of the stimulus and, more so, with attempts to conceal that relevance. The typical PDD examination is designed to elicit physiologic reactions from the examinee in response to questions concerning topic(s) of interest. Variability in Galvanic Skin Resistance (GSR), respiratory rate and/or volume, and heart rate/blood pressure are typically assessed (visually) by PDD examiners in the field. An increase in reactivity, defined as a change in response rate and/or amplitude, is interpreted as indicative of the examinee's truthfulness regarding the questions of interest.

Acknowledgments

The authors wish to express special thanks to: the volunteers who participated in the study; Charlene L. Stephens; Master Sergeant Randall S. Reynolds, United States Air Force; and the support staff at the Department of Defense Polygraph Institute (DoDPI), without whom this study could not have been completed. The authors would also like to thank Chief Warrant Officer Four Madison L. Mumbauer, United States Army, for his editorial assistance, and the Director of the DoDPI, Dr. William J. Yankee, for his continued support throughout the project. This manuscript is excerpted from a Federal Technical report of the same title. The full report is available from the Defense Technical Information Service # ADA299024. This study was supported by funds from the DoDPI as project DoDPI93-P-0025. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Editor's Note: This report is of a study conducted by the DoD Polygraph Institute in 1994. Since then, some PDD terminology has been updated. "GSR" has been superseded by electrodermal response (EDR), and the Control Question is now the Comparison Question. Because this article is excerpted from a DoDPI report, the original language is retained here.

Numerous valid criticisms have been expressed regarding the PDD process and associated assumptions (Furedy, 1986; Lykken, 1981; Office of Technology Assessment, 1983, pp. 29-43). Among those are the criticisms of validity and reliability of results. Validity is defined (Campbell, 1989, p. 749) as the degree to which a test measures what it is supposed to measure. Validity of a PDD examination would be measured as the degree of agreement between examiner decisions and ground truth (facts). Virtually all PDD studies attempt to assess the validity of PDD by comparing examiner decisions to ground truth. Definitions of ground truth range from experimental programming (i.e., asking subjects to participate in mock crimes so guilt and innocence are known quantities; Barland & Raskin, 1975) to decisions made by panels of experts who have reviewed case reports (Bersh, 1969). While questions of validity are very important, they are moot if reliable examination results are not obtained. Reliability is defined (Campbell, 1989, p. 629) as the degree to which a test measures the same thing consistently. A test of PDD examination reliability would require testing the same individual twice, using the same procedures. If PDD examinee responses are not consistent (among and/or between different measures), it is unlikely that questions of validity can ever be properly addressed. There have been numerous studies of interexaminer reliability in evaluating physiological data collected during a PDD examination (e.g., Horvath, 1977; Horvath & Reid, 1971; Hunter & Ash, 1973; Slowick & Buckley, 1975). Such studies are important in that they examine the consistency of data interpretation among examiners. These studies have not, however, investigated the reliability of physiologic responses.

Few studies report results concerning the consistency of examinee responses. An exploratory study completed by Ellson, Davis, Saltzman, and Burke (1952) was designed to examine the GSR responses of 10 male subjects using a variation of what is now labeled a stimulation card test (Abrams, 1989, pp. 120-122). They conclude that "one repetition of the detection procedure does not noticeably affect the success of the GSR as an indicator" of deception (Ellson et al., 1952, p. 7), but refer to no inferential statistics. Results of a second, similar, study confirm this hypothesis "unless the subject is told that the first attempt was successful" (Ellson et al., 1952, p. 11). Lieblich, Naftall, Shmueli, and Kugelmass (1974) employed a similar stimulation paradigm and GSR measure. They report that identification of deception was improved by repeating the same question sequence 10 times. Balloun and Holmes (1979) recorded the responses of 16 male subjects during two 5-question PDD examinations, separated by 30 seconds, administered using the Guilty Knowledge Questioning Technique (Lykken, 1960). They found that responses were attenuated during the second administration of the test and suggest that repeated examinations may be invalid. Grimsley and Yankee (1986) employed the Relevant/Irrelevant Question Technique to examine 80 male and female subjects on three occasions (separated by 24 hours). They found a non-significant decrease in accuracy between examinations 1 and 2, but no difference in accuracy between examinations 1 and 3. They conclude that overall accuracy rates are increased by evaluating multiple examinations. Yankee (1993) used the Control Question Technique (Reid, 1947) and a somewhat more realistic paradigm to investigate the accuracy of repeated examinations. Subjects (N = 72) were examined on two occasions, separated by 24 hours. Half of the subjects were programmed guilty via participation in a mock crime. Yankee also reported a decline in accuracy, though smaller in magnitude than that reported by Balloun and Holmes (1979), between the two examinations.

None of the investigations of repeated PDD examinations report data quantification beyond visual examination of physiological data. Decisions were usually based on visual inspection alone. Accurate absolute response levels are not mentioned. The very fundamental question of whether absolute response level differences occur during repeated examinations has not been investigated. The effect of a moderate delay between repeated examinations has also not been examined. The effect of such delays is simply not known and field examiners must rely on anecdotal knowledge for guidance. The current study is designed to examine response levels throughout repeated PDD examinations. A relatively simple variation of the Peak of Tension paradigm was chosen under the assumption that the results would generalize to more complex paradigms that use questions of greater personal relevance.

Method

Subjects

Forty-four, native English speaking, healthy males [mean age (SD) = 29.2 (7.8) years; range = 19 to 47] participated in this study. They were military personnel or Department of the Army civilian employees and were not paid for their participation. Thirty-nine of the subjects had never participated in a PDD examination before. The remaining five had not participated in a PDD examination within the last two years. Thirty-five of the subjects reported themselves to be medication free. The remainder had ingested pain/relaxant (3), anti-inflammatory (1), antibiotic (2), and antihistamine (3) medication within the 12-hour period prior to the examination. Females were not included because of possible variations in GSR (over-time) caused by hormonal secretions associated with the menstrual cycle.

Examiner

All PDD examinations were conducted by the same examiner, who had been trained at the United States Army Polygraph School and was certified by the United States Army as competent to administer PDD examinations. The examiner had administered approximately 500 field examinations during the 5 years prior to the study and was an instructor at the Department of Defense (DoD) Polygraph Institute at the time of the study. The examiner was not aware of whether subjects belonged to the control or treatment groups.

Apparatus

Data were collected using a Lafayette (Lafayette, IN) Factfinder (Model 76740/76741) polygraph equipped with three Cardio | Aux | Pneumo | GSR modules (Model 76477-G), one GSR module (Model 76480-G), and one electronic stimulus marker module (Model 76351-GET). Two of the multifunction modules (Model 76477-G) were used to record respiratory activity by setting the function selector to Pneumo and the third was used to record cardiovascular activity by setting the selector to Cardio-1. A circuit was added to the electronic stimulus marker module to allow control of the marker via signals from a computer RS-232 serial port. Lafayette sensors were used to measure GSR (Model 7664), respiration (Model 76513-1G & 76513- 2B), and cardiovascular activity (Model 76530).

An electronic circuit was designed and built in-house to amplify voltages from the Lafayette modules used to measure GSR, respiration, and cardiovascular activity. The amplified voltages were not affected by sensitivity or centering adjustments made to the instrument. Connection points for signal acquisition were: 1) GSR module - Pin 1, integrated circuit (I.C.) U1; 2) Cardio | Aux | Pneumo | GSR module - for respiration - Pin 3 of I.C. U1; and 3) Cardio | Aux | Pneumo | GSR module - for cardiovascular activity - Pin 7 of I.C. U2. The amplification circuit contained potentiometers that could be used to adjust the pre-amplifier voltage offset. A DC offset was indicated to be positive or negative by red/green LEDs mounted near the potentiometers. Amplification gains during testing were set at: 47x for the Pneumo/ respiration channels; 10x for the Cardio channel; and 5x for the GSR channel. Post-amplification signals were connected to a female 9-pin D connector. The amplification circuit module was inserted in an empty slot on the Lafayette polygraph and powered by the polygraph's internal power supply. The potentiometer controls, LED voltage indicators, and the 9-pin D connector were user accessible on the surface of the polygraph. Post-amplification physiologic signals were digitized using a Keithley Metrabyte (Taunton, MA) DAS-16F analog-to-digital converter mounted in an IBM PS/Value Point (Armonk, NY) Model 433DX microcomputer. Software was written in-house to digitize the physiologic signals at a rate of 256 samples/second.

A second micro-computer (Model 248, Zenith Data Systems, Chicago, IL), was used for question presentation. The questions used throughout testing were digitized and recorded to computer hard disk using a Sound Blaster board (Model 16ASP, Creative Labs Inc, Milpitas, CA). A parallel port interface (Speech Thing, Covox Inc., Eugene, OR), connected to a Radio Shack (Fort Worth, TX) integrated stereo amplifier (Model SA-155) and two speakers

(Model Minimus-77), was used to present the questions. This system ensured that each question was presented with the same inflection, and at the same volume, each time it was repeated. Computer software was written in-house to allow the examiner to present questions and digitize data by moving a cursor on the computer screen. Activation of the question presentation and data acquisition software was achieved via a serial port request-acknowledge algorithm.

Subjects' verbal responses were recorded on cassette tape (Tascam Model 134 4-channel recorder, TEAC, Montebello, CA) using a lavaliere microphone (Model 570S, Shure, Evanston, IL) held in place by a cord placed over the examinee's shoulders. The recorder was located in an adjacent room. Excerpt recording was controlled via the software running on the question presentation computer. The question presentation computer serial port and an in-house built interface for the cassette recorder were used for this purpose. Sound features of the audio recordings were extracted and examined as possible indexes of deception, as reported elsewhere (Cestaro & Dollins, 1994).

PDD testing was conducted in a carpeted, 3.5 x 3.66 m partially sound-attenuated room. Each examination was recorded on videotape using two ceiling and one wall-mounted video cameras. The examination was also monitored through a two-way mirror by a collaborator located in an adjacent room.

Subjects were seated in a Lafayette adjustable-arm subject chair (Model 76871, Lafayette, IN) during PDD testing. The chair was positioned beside and slightly in front of the examiner's desk. This position allowed the examiner to monitor the examinee's movements but not vice versa. The polygraph was mounted in a double pedestal examiner's desk (Lafayette Model # 76183). The question presentation and data acquisition computers and monitors were positioned on a table next to the examiner's desk and out of the examinee's sight during testing. The speakers, through which the questions were played, were located six feet behind, and one foot above, the back of the examinee's chair. The examinee's field of view, throughout testing, contained a wall of uniform color, a stationary video camera, and, above the video camera, a piece of paper with numbers and words written on it. Video cameras were also placed in the ceiling above the examiner's desk and behind the subject.

Procedure

Subjects were randomly assigned to the treatment or control groups, with the constraint that no more than three control or treatment group participants were tested consecutively. Twenty-two subjects were assigned to each group. Each subject participated in two examination sessions. The two examinations were separated by at least six working days. Subjects completed six Peak of Tension PDD tests during each examination session.

Upon arrival at the DoD Polygraph Institute (Fort McClellan, AL), each subject was escorted to a secluded briefing room and asked to read a brief description of the research project. Subjects who indicated that they would participate were asked to read and sign a volunteer agreement affidavit. Their questions were then answered. A brief biographical/medical questionnaire was then completed, to ensure that the subject was in good health and not currently taking medication which could interfere with the PDD examination results. The subject was then asked to complete a number search task, which was referred to as an anagram task. During this task, the participant circled six sequences of a two-digit number which was repeated five consecutive times (in any direction) in a 20 x 30 matrix of two digit numbers. The matrix consisted of numbers between 60 and 69 for the programmed deceptive subjects - who circled the number 64, and 80 to 89 for the programmed non-deceptive subjects - who circled the number 84. When the anagram task was completed, the subject was asked to write his name and the number he circled on two 7.62 x 12.7 cm cards. One card was retained by an investigator and the second concealed in the subject's pocket. The PDD examination procedure was briefly explained to the subject. It was emphasized that during the PDD examination the subject should not reveal which number he had circled when completing the anagram task. It was further emphasized that he should make every attempt to remain relaxed, even if he felt himself begin to react (increased heart rate, perspiration on hands, tightening of occlusive cuff) during the examination. The subject was then escorted to the examination room and introduced to the examiner.

The examiner greeted each subject, then reviewed the biographical/medical questionnaire with him to ensure it's accuracy. No other pretest questions were asked by the examiner. The examiner then briefly explained the sensors, procedures, and theory of PDD. The examiner explained that the polygraph measured physiological reactions - and not deception per se. It was further explained that the subject's physiological responses were likely to change during deception. It was suggested that fear of detection during deception altered the normal physiological response pattern and that these changes may be evident in the signals recorded during the PDD examination. The examiner described this response as being similar to the fight-or-flight reaction used to describe a fear response during military training. The examiner then reviewed the questions to be asked during data collection, with the subject, by playing the recorded questions.

All questions asked by the subject were then answered. He was then seated in the examination chair and the sensors were attached. Respiration was monitored using convoluted (pneumo) tubes placed around the thoracic area and abdomen. GSR was measured using stainless steel field electrodes placed, without paste, on the volar surface of the distal phalanges of the examinee's right hand index and ring fingers. Cardiovascular activity was monitored using an occlusive cuff placed over the brachial artery of the left arm. The pneumo tube vents were closed and the DC offsets for the pneumo and GSR were adjusted to zero. The sensitivity of these recording channels was then adjusted on the polygraph. Next, the occlusive cuff was inflated to 90 mmHg, massaged to remove wrinkles, then deflated to 48 mmHg. The pressure was then adjusted, as necessary, to achieve a 2 mmHg dial deflection between diastole and systole on the sphygmomanometer. The amplifier DC offset was then adjusted to zero, and polygraph sensitivity adjustments were made.

Each PDD test was composed of the following series of statements and questions, which were presented via computer recorded voice.

X The test is about to begin.
01 Did you complete an anagram for the number 60?
02 Did you complete an anagram for the number 61?
03 Did you complete an anagram for the number 62?
04 Did you complete an anagram for the number 63?
05 Did you complete an anagram for the number 64?
06 Did you complete an anagram for the number 65?
07 Did you complete an anagram for the number 66?
XX The test is now complete: please continue to sit still while I turn the instrument off.

If the examiner judged the physiological signals recorded on the polygraph chart to contain artifacts, the previous question was repeated. The examiner played the pre-recorded message "please remain still" if he judged that the examinee was producing unnecessary and/or excessive movements. When data collection for each test was completed, the pressure in the occlusive cuff was vented and the subject was instructed to "please relax while I prepare for the next test." If subjects appeared to be sleepy, they were also reminded of the importance of the study and encouraged to remain alert. The next PDD test was begun approximately three minutes later. The occlusive cuff was inflated, as described above, and DC offsets for the GSR and cardiovascular activity amplifiers were adjusted prior to beginning the next test. This process was repeated until six tests were completed, after which the sensors were removed. The subjects were then asked to read and sign a debriefing form, reminded to return the following week, and escorted out of the building.

Subjects returning for a second test session were escorted to a briefing room. They were reminded of the number circled during the previous session and asked to conceal the second card, indicating the number circled, in a pocket. They were reminded not to reveal the number they had circled to the examiner, then escorted to the examination room. The examiner again reviewed the biographical/ medical questionnaire from their previous session to ensure that no significant changes had occurred. Six additional PDD tests were completed, as described above. When the examination was completed, participants were thanked for their cooperation, asked to read and sign a second debriefing form, and escorted out of the building.

Data Reduction

The upper and lower pneumograph, GSR, and cardiovascular responses to each question were sampled at a rate of 256 samples per second for 14 seconds. Data sampling was initiated by the stimulus marker indicating that playback of the recorded question had ended. The data for each channel were smoothed to remove noise inherent in the instrument and/or amplifier used. Smoothing was implemented by substituting the average of the 50 points pre- and proceeding a data sample (i.e., a running average of 101 data points) for that sample. The first and last 50 data points of each epoch were then omitted from the epoch. This smoothing procedure was empirically determined to be the optimal solution to reducing noise in the recorded signal.

The data collected during day 1, test 3, questions 61 through 64 were lost, due to experimenter error, for 5 subjects (3 deceptive and 2 nondeceptive). Each response was reviewed for movement artifact contamination by three psychophysiologists who were blind to the treatment condition in which the sample was collected. Responses identified as containing movement artifacts by two or more reviewers were marked as missing data and omitted from further processing. All responses with amplitudes which exceeded the limits of the analog-to-digital converter were marked as missing data.

The following statistics were calculated for the remaining 13.6 second epochs. Line length of the upper and lower pneumograph (Pn1-LnL and Pn2-LnL, respectively), a technique introduced by Timm (1979, 1982a, 1982b), and GSR (GSR-LnL) data were calculated using a between point interval of 0.00390625 (i.e., 1/256). GSR peak amplitude (GSR-Amp) was calculated as the Peak Amplitude minus (0.5 * (Trough 1 + Trough 2) amplitudes). Troughs and peaks were identified as the first point where the subsequent 200 samples were greater (trough) or less (peak) than that point. If a peak was not identified within the first 7 seconds of data sampling, the peak amplitude values for the epoch were set to 0.000. Trough 1 was the first trough occurring prior to the peak or the first data sample if a peak but no trough was located. Trough 2 was the first trough identified after the peak. GSR peak latency (GSR-Ltc) was calculated, in seconds, relative to the first data point collected, for analysis where peaks were found. If a peak was not identified then the peak latency was considered missing data. The average heart rate inter-beat interval (CRD-IBI) epoch was calculated by determining the latency between the first and last R-wave peak found during the 13.6 second epoch and dividing by the total number of peaks found during the epoch - minus one.

The mean and standard deviation of responses recorded under each condition of the independent variables (group, day, test, and question) were calculated and only values within two standard deviations of the mean were retained for further analysis. (Note that data previously described as missing were omitted from this calculation.) All missing data were replaced by means from the appropriate condition combination. The proportion of missing data for each measure - by deceptive/non- deceptive group, respectively, was: Pn1-LnL - .07 / .07; Pn2- LnL - .07 / .09; GSR-LnL - .14 / .10; GSR-Amp - .15 / .12; GSR- Ltc - .25 / .20; and, CRD-IBI - .05 / .07.

It was observed that more than 50% of the GSR line length and amplitude data were missing for 2 subjects in each group and that more than 50% of the GSR peak latency data were missing for 6 subjects in each group. The data for these subjects were not analyzed for these measures.

Data Analysis

Statistical analyses were calculated using SYSTAT for DOS (Version 5.0) and

Windows (Version 5.04 - SYSTAT, Inc., Evanston, IL). The Pn1-LnL, Pn2-LnL, GSR-LnL, GSR-Amp, GSR-Ltc, and CRD-IBI response measures were initially analyzed using a between groups, within subjects 2(between-group) x 2(within-day) x 6(within-test) x 6(within-question) repeated measure analysis of variance (ANOVA). As mentioned above: 22 subjects per group were included in the Pn1-LnL, Pn2-LnL, and CRD-IBI analyses; 20 subjects per group were included in the GSR-LnL and GSR-Amp analyses; and 16 subjects per group were included in the GSR-Ltc analysis. A completely within subjects 2(day) x 6(test) x 6(question) repeated measure ANOVA was subsequently calculated, where appropriate, to resolve group main and interaction effects. The degrees of freedom used in calculating each mean square error term and F statistic were reduced by the proportion of missing data for that measure. F statistic probabilities of repeated measure effects with more than two levels were corrected for violations of sphericity assumptions using the Greenhouse - Geisser (1959) epsilon (ε). Orthogonal planned comparisons (Winer, 1971, pp. 172-215) were used to evaluate significant (p < 0.05) test and question main effects. The comparisons chosen to evaluate test effects were: (a) test 1 versus tests 2, 3, 4, 5, and 6; (b) test 2 versus tests 3, 4, 5, and 6; (c) test 3 versus 4, 5, and 6; (d) test 4 versus tests 5 and 6; and (e) test 5 versus test 6. Significant question effects were evaluated by comparing the measures recorded in response to questions concerning the numbers 62, 63, 64, 65, and 66 to those recorded in response to the remaining questions. For example, the responses following the question concerning the number 62 were compared to those concerning the numbers 61, 63, 64, 65, and 66.

The statistical power of each ANOVA F-test was calculated to assess the probability that the null hypothesis of no difference between the treatment means would be correctly rejected when the hypothesis was false (Williams & Zimmerman, 1989). Effect sizes were calculated as described by Cohen (1988, pp. 531-535), then converted to the noncentrality parameter, lambda, by multiplying the squared effect size by the number of observations in each analysis (Cohen, 1988, p. 550). It was necessary to convert effect sizes to a noncentrality parameter and calculate power directly rather than use Cohen's (1988) effect size directly because Cohen's (1988) tables underestimate the power of factorial designs (Koele, 1982). The denominator degrees of freedom used in the power calculations were reduced by the percent of missing data, as described above. Because the number of subjects in this design was relatively large, the power of each main effect and interaction was calculated using Laubscher's (1960, Formula 6) square root approximation of noncentral F (also described by Cohen, 1988, p. 550). The results of this approximation were cross-checked with Bavry's (1991, p. 127) calculation of the noncentral F distribution.

The power of the 2 x 2 x 6 x 6 ANOVA day x test, group x day x test, day x question, group x day x question, test x question, group x test x question, day x test x question, and group x day x test x question F-tests to detect an effect size of 0.20 was at least 0.80 - using a significance criterion of 0.05. The 2 x 2 x 6 x 6 ANOVA test, group x test, question, and group x question F- tests had a power of 0.80 to detect an effect size of 0.30 using a significance criterion of 0.05. The 2 x 2 x 6 x 6 ANOVA had relatively low power to detect group, test, and group x test effect sizes due to the small number of observations in these analyses. The power of reported statistical differences was at least 0.80 at a critical significance level of 0.05 or less. The degrees of freedom used during power calculation were adjusted to compensate for possible violation of sphericity assumptions using the Greenhouse and Geisser epsilon (Geisser & Greenhouse, 1958; Greenhouse & Geisser, 1959; Winer, 1971, p. 523), as suggested by Keppel (1991, pp. 355-356).

Results

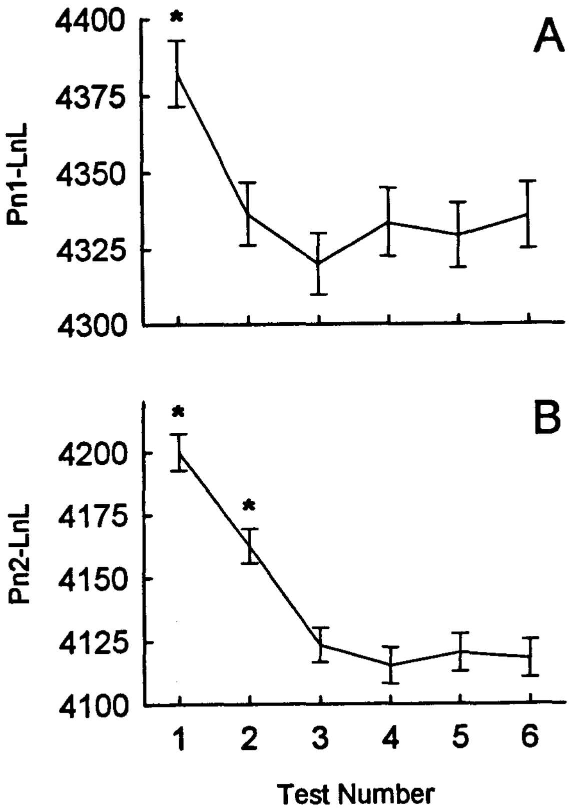
Pn1-LnL (Upper Pneumograph Line Length)

Pn1-LnL changed significantly over repeated tests [F(5, 195) = 3.35, ε = .70, p = .015]. Planned comparison results indicated that the average Pn1-LnL measured during test 1 was 51.27 (Average SEM = 10.67) longer [F(1, 39) = 9.981, p = .003] than the average of those measured during tests 2, 3, 4, 5, and 6. This difference is illustrated in Figure 1-A.

The group x question interaction was also significant [F(5, 195) = 2.84, ε = .60, p = .041].

The deceptive and nondeceptive subject responses were analyzed separately to facilitate interpretation of the group x question interaction (Keppel, 1991, pp. 383-384). A significant question effect [F(5, 98) = 3.59, ε = .39, p = .038] was found among the deceptive subject responses, but not among those of the nondeceptive subjects. The results of subsequent comparisons among deceptive subject responses to questions, illustrated in Figure 2-A, indicated that the average response to the question concerning the number 64 was 47.19 (Average SEM = 14.23) shorter [F(1, 20) = 17.13, p = .000] than the average of the remaining question responses.

Figure 1. Mean (SEM) response levels for A) Pn1-LnL and B) Pn2-LnL averaged over questions, days, and groups. Values marked with an asterisk (*) are significantly greater than subsequent values.

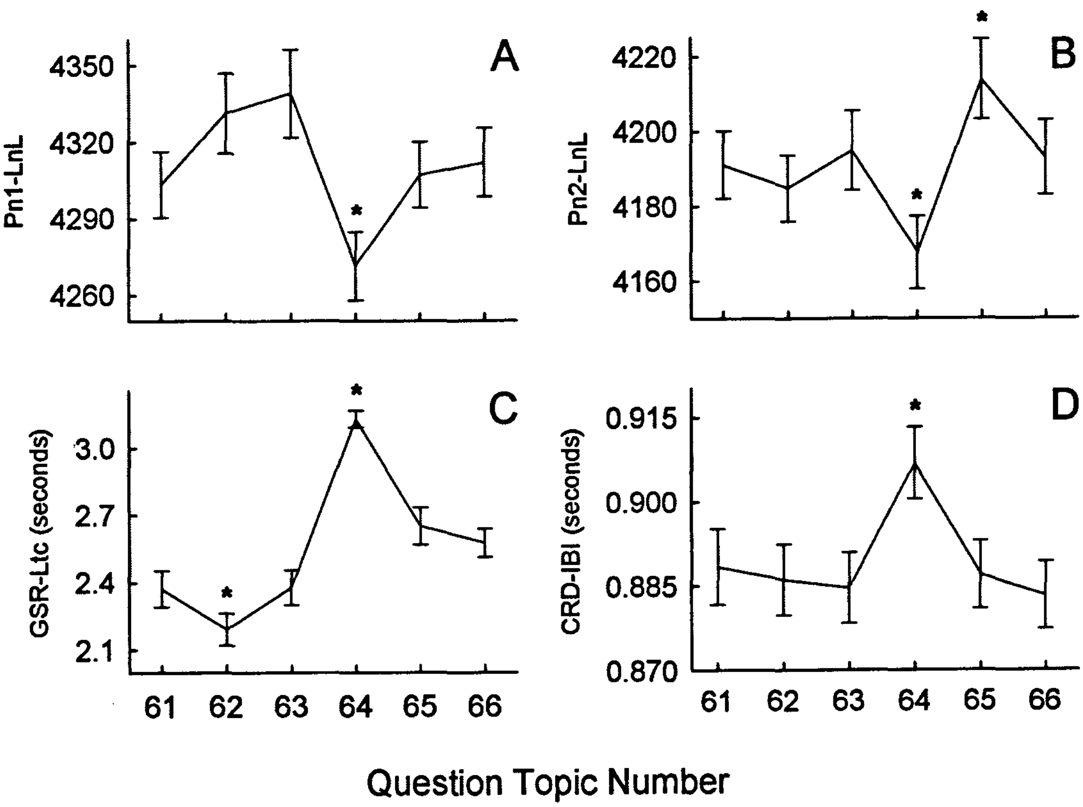


Pn2-LnL (Lower Pneumograph Line Length)

Pn2-LnL responses measured from the deceptive subjects were an average of 101.76 (Average SEM = 4.03) longer than those measured from nondeceptive subjects [F(1, 39) = 9.40, p = .004]. Pn2-LnL also changed significantly over repeated tests [F(5, 193) = 14.89, ε = .83, p = .000]. Results of planned comparisons indicated that the average Pn2-LnL measured during test 1 was 80.82 (Average SEM = 7.19) longer [F(1, 39) = 46.03, p = .000] than the average Pn2-LnL of subsequent tests, and that the average Pn2-LnL measured during test 2 was 72.16 (Average SEM = 7.19) longer [F(1, 39) = 18.02, p = .000] than the average measured during tests 3, 4, 5, and 6, as illustrated in Figure 1-B. While a significant question effect was found [F(5, 193) = 3.76, ε = .82, p = .005], the planned contrasts were all non- significant. The group x question interaction was also significant [F(5, 193) = 5.07, ε = .82, p = .000].

The deceptive and nondeceptive subject responses were analyzed separately to facilitate interpretation of the group x question interaction. A result of these analyses was that responses were shown to change significantly over repeated tests for both groups. The results of subsequent comparisons among tests showed the same pattern of significant effects as the overall analysis. Responses measured from the deceptive subjects differed significantly during question repetition [F(5, 97) = 5.52, ε = .66, p = .000], while those measured from the nondeceptive subjects did not. Comparison results, illustrated in Figure 2-B, indicate that the deceptive subjects' average Pn2-LnL response to the question concerning the number 64 was 27.96 (Average SEM = 9.82) shorter [F(1, 19) = 9.05, p = .007] than those in response to the remaining questions. In addition, the deceptive subjects' average Pn2-LnL response to the question concerning the number 65 was 27.43 (Average SEM = 9.82) longer [F(1, 19) = 11.04, p = .004] than the average Pn2-LnL responses to the remaining questions. Responses measured from non-deceptive subjects were also found to differ significantly during question repetition [F(5, 95) = 3.09, ε = .65, p = .030], but no significant differences were found among the subsequent planned comparisons.

Figure 2. Deceptive subjects' mean (SEM) response levels for A) Pn1-LnL, B) Pn2-LnL, C) GSR-Ltc, and D) CRD-IBI averaged over tests and days. Values marked with an asterisk (*) are significantly greater or less than the average of the remaining values.



GSR-LnL (Galvanic Skin Response Line Length)

A significant group x day x chart interaction [F(5, 167) = 3.49, ε = .86, p = .007] was found among the GSR-LnL measures, but simple effect analysis did not reveal where the differences occurred.

GSR-Amp (Galvanic Skin Response Amplitude)

The average GSR-Amp measured from the deceptive subjects was 10.56 (Average SEM = .877) greater [F(1, 33) = 10.35, p = .002] than that measured from the nondeceptive subjects. Average GSR-Amp responses also changed significantly [F(5, 165) = 3.21, ε = .85, p < .013] among repeated tests. Planned comparisons, however, failed to reveal any significant differences. Significant group x question [F(5, 165) = 13.29, ε = .79, p = .000] and group x day x chart [F(5, 165) = 3.49, ε = .84, p = .008] interactions were also found.

Separate analyses of the deceptive and nondeceptive subject GSR-Amp responses were calculated to facilitate interpretation of the group x question and group x day x chart interactions. A significant difference was found among the question responses of the nondeceptive subjects [F(3, 83) = 9.71, ε = .50, p = .000]. Planned comparisons indicated that the average GSR-Amp recorded in response to the question concerning the number 62 was 9.50 (Average SEM = 1.70) greater than the average GSR-Amp response to the remaining questions [F(1, 16) = 11.34, p = .004]. The average GSR-Amp recorded in response to the question concerning the number 63 was 5.37 (Average SEM = 1.70) less than the average response to the remaining questions [F(1, 16) = 13.51, p = .002]. Significant differences were also found among the average question [F(5, 80) = 6.92, ε = .74, p = .000] and test [F(5, 80) = 2.81, ε = .74, p = .021] responses of the deceptive subjects. The deceptive subject average GSR-Amp response (Average SEM = 2.39) to the question concerning the number: 62 was 12.73 smaller [F(1, 16) = 22.25, p = .000] than that to the remaining questions; and, 66 was 10.46 smaller [F(1, 16) = 16.79, p = .000] than that to the remaining questions. No significant differences were found among the planned comparisons for the tests.

GSR-Ltc (Galvanic Skin Response - Response Latency)

A significant GSR-Ltc measure difference was found among responses to the questions asked during testing [F(5, 115) = 9.29, ε = .84, p = .000]. Comparisons indicate that response latencies to the question concerning the number 63 were 0.23 seconds (Average SEM = .057) shorter [F(1, 23) = 11.93, p = .002] than those to the remaining questions. Response latencies to the question concerning the number 64 were 0.43 seconds (Average SEM = .057) longer [F(1, 23) = 49.33, p = .000] than the average of those recorded in response to questions concerning the numbers 61, 62, 63, 65, and 66. The 2 x 2 x 6 x 6 ANOVA also indicated that there was a significant group x question effect [F(5, 115) = 8.62, ε = .84, p = .000].

Data recorded from the deceptive and nondeceptive groups were analyzed separately to assist in interpreting the significant group x question effect. Significant question effects were found for both the nondeceptive [F(5, 60) = 5.01, ε = .65, p = .001] and deceptive [F(5, 56) = 19.69, ε = .76, p = .000] subject responses. No significant differences were found among the question effect planned comparisons for the nondeceptive group. The average deceptive subject GSR-Ltc response (Average SEM = .069) to the question concerning the number 62 was 0.43 seconds shorter [F(1, 11) = 33.75, p = .000] than the average response to the remaining questions. The average deceptive subject GSR-Ltc response to the question concerning the number 64 was 0.69 seconds longer [F(1, 11) = 105.44, p < .000] than the average response to the remaining questions. These differences are illustrated in Figure 2-C.

A significant day x test x question effect [F(25, 281) = 2.88, ε = .35, p = .000] was found among the responses of the deceptive subjects. Separate analyses were calculated for the deceptive subject responses recorded during test days 1 and 2 to assist in interpreting this effect. These analyses indicated significant differences among the GSR-Ltc question responses for both day 1 [F(5, 56) = 6.07, ε = .71, p = .001] and day 2 [F(5, 56) = 10.20, ε = .68, p = .000]. Planned comparisons indicated that the deceptive subject GSR-Ltc responses to the question concerning the number 64, during day 1, were .65 (Average SEM = .10) seconds longer [F(1, 11) = 41.77, p = .000] than the average response latency to the remaining questions. Comparisons for deceptive responses measured during day 2 (Average SEM = .09) indicate that responses to the question 62 were .38 seconds shorter [F(1, 11) = 59.36, p = .000] than the average response latency to the remaining questions and that responses to the question concerning the number 64 were .72 seconds longer [F(1, 11) = 41.37, p = .000] than the average response latency to the remaining questions.

A significant test x question effect was found among the deceptive subjects GSR-Ltc responses during test day 2 [F(25, 281) = 2.22, ε = .32, p = .033]. Each test was analyzed separately to assist in interpreting this difference. No significant differences were found among the question responses recorded during tests 1 and 5. The analyses indicated that there were significant differences among responses recorded to questions during tests 2, 3, 4, and 6. Contrasts indicate that the average GSR-Ltc responses to the question concerning the number 64 were significantly longer (p < .05) than the average of those recorded in response to the remaining questions during tests 2, 3, 4, and 6. GSR-Ltc responses to questions concerning the numbers 62 and 66 recorded during test 4 and to questions concerning the number 62 recorded during test 6 were significantly shorter (p < .05) than the average of the responses recorded during the remaining questions.

Separate analyses were calculated for the nondeceptive subject responses recorded during test days 1 and 2 to assist in interpreting a significant day x test effect result found during the analysis of nondeceptive subject GSR-Ltc responses [F(5, 60) = 2.72, ε = .68, p = .050]. No significant test, question, or test x question effects were found among the non-deceptive subject GSR-Ltc responses recorded during day 1. Nondeceptive subject responses on day 2 were, however, found to differ significantly among questions [F(5, 60) = 4.46, ε = .623, p = .002]. Planned comparisons indicate that the average GSR-Ltc response latency to the question concerning the number 63 was .37 (Average SEM = .12) seconds shorter than the average response latency to the remaining questions [F(1, 12) = 13.71, p = .003].

CRD-IBI (Cardio Channel Average Inter-beat-interval)

A significant CRD-IBI measure difference was found among responses to the questions asked during testing [F(5, 197) = 4.27, ε = .53, p = .009]. Comparisons indicate that the average CRD-IBI measured in response to the question concerning the number 64 was 0.011 (Average SEM = .005) seconds longer [F(1, 39) = 14.80, p = .000] than those to the remaining questions. The 2 x 2 x 6 x 6 analysis also indicated significant group x question [F(5, 197) = 3.41, ε = .53, p = .025], group x day x test [F(5, 197) = 3.06, ε = .83, p = .017], and group x test x question interactions [F(25, 987) = 1.93, ε = .45, p = .03].

Separate analyses of the data recorded from the deceptive and nondeceptive subjects were calculated to facilitate interpretation of the significant interaction effects. The analysis indicated no significant differences among the nondeceptive subject responses as a function of the independent variables manipulated. A significant question effect was, however, found among the deceptive subject responses [F(5, 99) = 5.84, ε = .54, p = .002]. Planned comparisons indicated that the deceptive subjects' average CRD-IBI response to the question concerning the number 64 was .021 (Average SEM = .0061) seconds longer than the average response CRD-IBI to the remaining questions, as illustrated in Figure 2-D.

Discussion

Interpretation of these results suggests that during repeated administration of PDD tests: there is a consistent change in average Pn1-LnL and Pn2-LnL; differential Pn1-LnL, Pn2-LnL, and CRD-IBI reactivity during a PDD test does not change during repeated tests or days; and, average physiological reactivity of deceptive subjects changes during deception while that of nondeceptive subjects does not. When interpreting these results it is important to remember that the power of each significant statistical effect was 0.80 or greater and that the power of the non-significant statistical tests to detect an effect of size 0.30 at the 0.05

significance level was also 0.80 or greater (with exceptions noted above). The power analysis provides the probability (0.80 or greater) that the null hypothesis is correctly rejected when a significant effect was observed, as well as the probability (0.80 or greater) that an effect size of 0.30 would have been correctly detected.

Perhaps the most interesting result of this research is not the significant results which were obtained, but those that were not. All day x test, day x question, test x question, and day x test x question interactions were non-significant. This suggests that the pattern and/or variability of measured physiologic responses to the questions asked during each PDD test did not change significantly over repeated administration of the tests, nor did the response pattern change significantly between days 1 and 2 - with the exception of GSR-Ltc responses. This result is interpreted as supporting those of Ellson et al. (1952), Lieblich et al. (1974), Grimsley and Yankee (1986), and Yankee (1993) that there were no statistically significant differences in the detection of veracity with repeated testing. While veracity detection rates were not determined, the conclusion that differential responding does not change with question series repetition supports the proposal that decision accuracy does not decrease with repeated testing (Grimsley & Yankee, 1986; Iacono, Boisvenu, & Fleming, 1984; Leiblich et al., 1974).

The results of some investigations into the effect of repeated question series administration on skin resistance and/ or conductance responsivity do not support those of this study (Balloun & Holmes, 1979; Ben-Shakhar & Lieblich, 1982; Elaad & Ben-Shakhar, 1989; Iacono et al., 1984) while those of others do (Furedy & Ben-Shakhar, 1991; Furedy, Gigliotti, & Ben-Shakhar, 1994). This is a difficult issue to resolve due to methodological differences in the: response requirements; question repetition patterns and procedures; and, data reduction, evaluation, and analysis techniques. It is also possible that the response strengths measured during this study decreased with repetition, but the decrease was too small to be statistically detected. It is likely, however, that such small changes would be of little interest. Further research should be conducted to address these issues.

Average Pn1-LnL and Pn2-LnL response levels measured during the first test, averaged over groups, days, and questions, were found to be significantly greater than the average of the subsequent tests, as illustrated in Figure 1. No statistically significant difference was found between Pn1-LnL measures recorded during tests 2 through 5 and the average of subsequent tests. The average Pn2-LnL measure recorded during test 2 was significantly greater than the average recorded during tests 3 through 6, but measures recorded during tests 3 through 5 were no different from those recorded during subsequent tests. A similar shift in skin conductance following repeated testing has been reported by Iacono et al. (1984). The decrease in average response levels observed during the initial stages of repeated testing, in the absence of within test response attenuation, may be a variation of the phenomenon of differential autonomic responsivity, proposed by Ben-Shakhar and Lieblich (1982).

Results of the data analyses indicate that there were no statistically significant main or interaction effects related to the questions asked among the average nondeceptive subject Pn1- LnL, Pn2-LnL, and CRD-IBI responses. The average deceptive subjects' deceptive responses were shorter in Pn1-LnL and Pn2-LnL, longer in GSR-Ltc, and longer in CRD-IBI than the average of their nondeceptive responses. These results confirm that, on the average, a pattern of differential responding occurs during deception that does not occur when deception is not present. While pneumo line lengths and heart rate are not normally evaluated when scoring PDD examinations, perhaps polygraphs used for PDD should be modified to display this information.

While significant differences were found among the deceptive subjects' GSR-Amp responses to the questions asked, the deceptive response was not significantly different from the average nondeceptive response. This is surprising when one considers results of studies reporting high veracity detection accuracy rates based exclusively on electrodermal activity scores (Iacono, Cerri, Patrick, & Fleming, 1992; Kugelmass & Lieblich, 1966; Podlesny & Raskin, 1978; Thackary & Orne, 1968). However, close examination of these reports suggests that differences in methodology and evaluation techniques could account for the differences between the current results and earlier reports. While a field polygraph was used in the current study, the operator sensitivity adjustments were bypassed. Skin resistance changes were amplified by a fixed-gain linear amplifier adjusted to remain within the range limits of an analog-to-digital converter, which did not compensate for changes in tonic skin resistance, possibly contributing to the failure to find significant differences among GSR-Amp measures during deception, in this study.

It should, however, be noted that 9% and 27% of the subjects were dropped from the GSR-Amp and GSR-Ltc analyses, respectively, due to insufficient data caused, primarily, by failure to obtain quantifiable subject responses. The percentages of missing Pn1-LnL, Pn2-LnL, and CRD-IBI data, which were collected simultaneously with the GSR data, were not sufficiently large to necessitate removal of subjects from the analyses. This observation is interpreted as suggesting that the exclusive or disproportionately high reliance on GSR response scores when interpreting the results of PDD examinations may lead to excessive errors. This suggestion is not new, but simply reinforces the statement presented to the Committee on Government Operations over 20 years ago that "most examiners agree that the galvanic skin response is the least accurate, and should be ignored when a conflict (among the three channels) occurs" (Committee on Government Operations, 1974, p. 24).

In summary, three conclusions are derived from the results of this research. First, a consistent change was observed in average Pn1-LnL and Pn2-LnL responses, but not the GSR-Amp, GSR- LnL, GSR-Ltc, and CRC-IBI responses as the test was repeated. This pattern did not change significantly between test days one and two. Second, the average physiological response variability measured during a PDD test did not change over repeated tests. Finally, the Pn1-LnL, Pn2-LnL, GSR-Ltc, and CRD-IBI responses of deceptive subjects, averaged over repeated test administrations, changed during the deceptive response, relative to nondeceptive responses. No such systematic changes were found among the responses of the nondeceptive subjects. These data are interpreted as suggesting that decision accuracy will not decrease significantly during repeated (up to six) administrations of the question series during a PDD examination. This conclusion is supported by other reports (Grimsley & Yankee, 1986; Iacono et al., 1984; Leiblich et al., 1974). It is suggested that changes in heart rate inter-beat-interval, measured using an occlusive cuff as described, and pneumo line length are reliable response measures which may be accurately interpreted as indicating deception.

References

Abrams, S. (1989). *The complete polygraph handbook.* Lexington, MA: Lexington Books.

Balloun, K. D., & Holmes, D. S. (1979). Effects of repeated examinations on the ability to detect guilt with a polygraphic examination: A laboratory experiment with a real crime. *Journal of Applied Psychology*, 64, 316-322.

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

Bavry, J. L. (1991). *Stat-Power statistical design analysis system.* Chicago, IL: Scientific Software, Inc.

Ben-Shakhar, G., & Lieblich, I. (1982). The dichotomization theory for differential autonomic responsivity reconsidered. *Psychophysiology,* 19, 277-281.

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

Campbell, R. J. (1989). *Psychiatric Dictionary (6th ed).* New York: Oxford University Press.

Cestaro, V. L., & Dollins, A. B. (1994). An analysis of voice responses for the detection of deception (Report No. DoDPI94- R-0001). Ft. McClellan, AL: Department of Defense Polygraph Institute.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Committee on Government Operations (1974). The use of polygraphs and similar devices by federal agencies (Hearings before a Subcommittee of the Committee on Government Operations, House of Representatives, 93rd Congress, 2nd Session). Washington, DC: U.S. Government Printing Office.

Elaad, E., & Ben-Shakhar, G. (1989). Effects of motivation and verbal response type on psychophysiological detection of information. *Psychophysiology,* 26, 442-451.

Ellson, D. G., Davis, R. C., Saltzman, I. J., & Burke, C. J. (1952). Accuracy of detection and the effect of repetition. A report of research on detection of deception (Tech. Rep. Contract No. N60NR-18011). Bloomington, IN: Indiana University, Department of Psychology.

Furedy, J. J. (1986). Lie detection as a psychophysiological differentiation: Some fine lines. In M. G. H. Coles, E. Donchin, & S. W. Porges (Eds.), *Psychophysiology: Systems, Processes, and Applications* (pp. 683-701). New York: Guilford Press.

Furedy, J. J., & Ben-Shakhar, G. (1991). The roles of deception, intention to deceive, and motivation to avoid detection in the psychophysiological detection of guilty knowledge. *Psychophysiology,* 28, 163-171.

Furedy, J. J., Gigliotti, F., & Ben-Shakhar, G. (1994). Electrodermal differentiation in deception: The effect of choice versus no choice of deceptive items. *International Journal of Psychophysiology,* 18, 13-22.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics,* 29, 885-891.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika,* 24, 95-112.

Grimsley, D. L., & Yankee, W. J. (1986). The effect of multiple retests on examiner decisions in applicant screening polygraph examinations (National Security Agency Contract No. MDA 904-85-C-A962). Charlotte, NC: A. Madley Corporation.

Horvath, F., & Reid, J. (1971). The reliability of polygraph examiner diagnosis of truth and deception. *Journal of Criminal Law, Criminology, and Police Science,* 62(2), 276-281.

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

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

Iacono, W. G., Boisvenu, G. A., & Fleming, J. A. (1984). Effects of diazepam and methylphenidate on the electrodermal detection of guilty knowledge. *Journal of Applied Psychology*, 69, 289-299.

Iacono, W. G., Cerri, A. M., Patrick, C. J., & Fleming, J. A. E. (1992). Use of antianxiety drugs as countermeasures in the detection of guilty knowledge. *Journal of Applied Psychology*, 77, 60-64.

Keppel, G. (1991). *Design and analysis, a researcher's handbook (3rd ed.).* Englewood Cliffs, NJ: Prentice Hall.

Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin*, 92, 513-516.

Kugelmass, S., & Lieblich, I. (1966). The effects of realistic stress and procedural interference in experimental lie detections. *Journal of Applied Psychology*, 50, 211-216.

Laubscher, N. F. (1960). Normalizing the noncentral t and F distributions. *Annals of Mathematical Statistics*, 31, 1105- 1112.

Lieblich, I., Naftali, G., Shumueli, J., & Kugelmass, S. (1974). Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. *Journal of Applied Psychology*, 59, 113-115.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, 44, 258-262.

Lykken, D. T. (1981). *A Tremor in the Blood: Uses and Abuses of the Lie Detector.* New York: McGraw-Hill.

Office of Technology Assessment (1983). Scientific Validity of Polygraph Testing: A Research Review and Evaluation - A Technical Memorandum (Report No. OTA-TM-H-15, November). Washington, DC: Office of Technology Assessment.

Podlesny, J., & Raskin, D. (1978). Effectiveness of techniques and physiological measures in detection of deception. *Psychophysiology*, 15, 344-359.

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law & Criminology*, 37, 542-547.

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

Thackray, R., & Orne, M. (1968). A comparison of physiological indices in detection of deception. *Psychophysiology*, 4, 329- 339.

Timm, H. (1979). The effect of placebos and feedback on the detection of deception (U.S. Department of Justice Contract No. 78-NI-AX-0028). East Lansing, MI: Michigan State University.

Timm, H. (1982a). Analyzing deception from respiration patterns. *Journal of Police Science and Administration*, 10(1), 47-51.

Timm, H. (1982b). Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. *Journal of Applied Psychology*, 67(4), 391-400.

Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology*, 116, 359-369.

Winer, B. J. (1971). *Statistical principles in experimental design (2nd ed.).* New York: McGraw-Hill.

Yankee, W. J. (1993). Test and re-test accuracy with a psychophysiological detection of deception test. Department of Defense Polygraph Institute, Ft. McClellan, AL. Adopted from Yankee, W. J. & Grimsley, D. L. (1986) The effect of a prior polygraph test on a subsequent polygraph test (NSA contract No. MDA 904-84-C-4249). Charlotte, NC: University of North Carolina.

A Brief Note on the Misleading and the Inaccurate: A Rejoinder to Matte (2000) With Critical Comments on Matte and Reuss (1999)

Charles R. Honts

Abstract

Matte (2000) is self-described as a critical analysis of Honts (1999). Rather than being a critical analysis of Honts (1999) Matte (2000) is instead an inaccurate, misleading, and often ad hominem polemic. The present paper sets the record straight by correcting Matte mistakes and misrepresentations. Nothing in Matte (2000) weakens the arguments presented in Honts (1999). All of the available data, from real examinations, support the review of questions between repetitions.

Key words: inter-chart stimulation, reply

Matte (2000) was self-described as a "critical analysis" of Honts (1999) paper on the review of comparison questions between chart representations. Unfortunately, Matte (2000) is full of misrepresentations, misguided statements and inaccuracies. Matte (2000) was published contiguously with Honts, Raskin, Amato, Gordon, & Devitt (2000) in the same issue of this journal. Honts et al., (2000) responded to earlier comments by Abrams (1999). Much of the material in Honts et al., (2000) speaks directly to Matte's (2000) comments and clearly shows his conclusions to be incorrect.

The Hypothetical Construct, Psychological Set

Matte (2000) begins with a number of unattributed statements regarding an unspecified hypothetical construct, psychological set, and a number of hypothesized impacts of discussion of questions between repetitions on psychological set. The notion of psychological set is a contrivance of the polygraph profession and has received little scientific validation. Moreover, psychological set is not a term that is currently much used in mainstream psychological science. While the hypothetical construct, psychological set, may have some heuristic value as a descriptive tool, it has no reality in science or the real world. Nevertheless, Matte asserts as fact that the discussion of relevant questions between repetitions would result in an increased false positive rate, while a discussion of comparison questions between charts would result in an increase false negative rate. Matte cites no data and no studies to support these alleged facts. I am aware of no data that have tested the former proposition, but data exist that directly test the latter. Those data were reproduced in Table 1 of Honts et al., (2000). The two studies (Dawson, 1981; Patrick and Iacono, 1989) described there were published in first tier, peer-reviewed scientific journals. Those two studies directly tested Matte's (2000) hypothesis about increased false positives following between chart stimulation of only comparison questions, and showed his speculation to be incorrect. In the Dawson study, only comparison questions were stimulated between charts and 100% of the decisions on guilty subjects were correct. A similar, but less dramatic pattern of results were reported by Patrick and Iacono (1989). Matte must be aware of these data but refuses to acknowledge that they prove his hypothesis about the review of comparison questions to

Author Notes

Dr. Honts is a Professor and Chair of the Department of Psychology at Boise State University. In addition he has been a practicing polygraph examiner since 1976. He has published and/or presented over 200 scientific papers, most of them on the topic of the detection of deception. Please address correspondence to: Charles R. Honts, Ph. D., Department of Psychology, Boise State University, 1910 University Drive, Boise, ID 83725-1715. E-mail: chonts@boisestate.edu

be false. Moreover, they raise serious questions about the validity of his entire theory of the functioning of comparison question tests.

Matte & Reuss (1999): A Study Without External Validity

Matte then uses data from Matte and Reuss (1999) in an effort to say that directed-lie comparison questions are somehow uniquely susceptible to deleterious impacts on false negative rates from review between question repetitions. Nothing could be further from the truth. In a game paradigm, Matte and Reuss had subjects imagine that they had committed a crime and then they had their subjects pretend that they were taking a polygraph examination. Furthermore, the written game scenario given to the subjects as the stimulus material presents a very unrealistic picture of taking a polygraph examination. Essentially no discussion of the relevant issue is mentioned, with only 82 words being devoted to the entire section on the issues of the examination and the relevant questions. The comparison questions are given 247 words, three times as many as was the relevant issue and the relevant questions. This is an utterly unrealistic depiction of a real polygraph examination where an hour or more may have been devoted to the discussion of the relevant issues and the development of the relevant questions while only a few minutes would have been spent presenting the comparison questions. Then, Matte and Reuss' subjects were then asked to imagine how much threat they felt from the various questions posed while they played this contrived game. None of their subjects ever enacted a crime, took a polygraph examination, nor received any reward or punishment associated with their participation. Matte and Reuss reported that 86% of the people who pretended they were guilty as part of this game reported that they would find the directed-lie comparison questions to be of equal or greater threat than the relevant questions. Given the unrealistic methods used in this game, the results are not at all surprising, nor are they informative about any real-world practices.

Mock crime laboratory studies are sometimes criticized for lacking generalizability (external validity) to real-world polygraph examinations. This, despite the fact that in the better laboratory studies subjects enact a mock theft, are given polygraph examinations by real polygraph examiners, with real polygraphs, and have rewards and sometimes punishments associated with the outcomes of those examinations. If such high quality mock crime laboratory studies are considered by some (e.g. Iacono & Lykken, 1997) to be weak in external validity, then Matte and Reuss (1999) must be viewed as having no external validity whatsoever. Moreover, even if one were to credit Matte and Reuss with some connection to the real world, their data would still have no meaning for the actual physiological data collected in polygraph examinations. Human beings are notoriously poor perceptors of their physiological responses. In studies of the subjects' posttest perceptions of the importance of test questions and their physiological responses to them, no relationship has been found between their perceptions, their guilt status, or their actual physiological responses (Honts, 1986; Horowitz, Kircher, Honts & Raskin, 1997). Matte and Reuss (1999) tells us nothing about how individuals respond physiologically in a real directed-lie polygraph examination or in any other real world situation. For a discussion of data concerning the directed lie comparison test, readers are referred to Honts and Gordon (1999), Honts et al., (2000) and to Raskin, Honts, and Kircher (1997).

Case Selection, Support for, Rather Than a Criticism of Honts (1999)

Matte (2000) criticizes Honts (1999) for including studies that contain different methodologies. While it is true that the studies reported in Honts (1999) do contain different methodologies, this is not a valid criticism of the analysis reported therein. Matte (2000) betrays a serious lack of knowledge of scientific and research methodology in that the collective study of other studies is an accepted and well-known scientific approach known as meta-analysis (Hunter, Schmidt & Jackson, (1982). Moreover, meta-analysis is a long accepted way for summarizing the results of many studies of polygraph examinations and for examining variables across the various studies (e.g. Kircher, Horowitz, & Raskin, 1988; Office of Technology Assessment, 1983). The fact that I reported large effects of a reduction of false negative errors due to the discussion of questions between question repetitions (Honts, 1999), despite the fact that the studies used disparate methodologies, is powerful evidence for the strength of my arguments, not a criticism of them.

Matte (2000) then criticizes Honts (1999) for including the Szucko & Kleinmuntz (1981) study. The only cited reason for questioning the inclusion of Szucko and Kleinmuntz is reference to an unpublished personal communication that allegedly occurred between one of the participating examiners and Frank Horvath. The Szucko and Kleinmuntz study was published in the journal *Nature*. *Nature* is one of the premier multi-disciplinary scientific journals in the world. It is peer-reviewed and rejects something on the order of 90% of the articles submitted to it for consideration of publication. Matte's suggestion that we should reject such a publication on the basis of an unsubstantiated personal communication is, at best, sophistry. While I and others have stated objections to the use of Szucko and Kleinmuntz as an estimator of the accuracy of polygraph tests in general (Raskin et al., 1997), nothing about the design of Szucko and Kleinmuntz should have impacted the relationship being studied in Honts (1999). The study met the criterion specified in the method of Honts (1999), to have left the study out of the meta-analysis simply because one did not like the results would have shown true bias and poor methodology. It is peculiar that Matte (2000) picked the Szucko and Kleinmuntz study for specific criticism. Performance in Szucko and Kleinmuntz with guilty subjects was above the median. Elimination of Szucko and Kleinmuntz from Honts (1999) would strengthen the Honts (1999) argument and weaken Matte's (2000) position.

Matte (2000) then criticizes Honts (1999) for its method of study selection, the implicit suggestion is that the study selection was biased. Matte quotes (without correct attribution), Honts (1999) as follows:

> "The studies shown in Table 1 were selected for inclusion in the analysis because they met at least one of the following criterion: The method section of the study explicitly described the discussion of, or the lack of discussion of, comparison and/or relevant questions between question list repetitions."

However, the complete quotation is as follows:

> The polygraph literature available to the author was reviewed for information concerning the review of comparison questions between question list repetitions. The studies shown in Table 1 were selected for inclusion in the analysis because they met at least one of the following criterion:
>
> * The method section of the study explicitly described the discussion of, or the lack of discussion of, comparison and or relevant questions between question list repetitions.
>
> * The present author was involved in the conduct of the study and had a personal knowledge of the methods used.
>
> * The study was conducted within an organization that has an explicit policy regarding the review of comparison questions between question list repetitions.
>
> Studies that were not classifiable under one of the three criteria were not included in the analysis. (Honts, 1999, p.119)

Based upon this misrepresentation of my selection criteria, Matte suggests that "many other studies which reflected significantly higher accuracy rates could have qualified for inclusion..." (p. 147). Given that he has misrepresented my selection methods, Matte's suggestion that I engaged in selective scholarship is sophistry at best, and purely disingenuous at worst. Moreover, he fails to name even one study that he thinks should have been included. While it might be legitimate to make an argument that the Honts (1999) selection criteria were flawed, such an argument should include other suggested criteria and a list of studies whose exclusion were clearly of a biased nature. Matte (2000) offers neither.

Ad Hominem Attacks

Finally, Matte (2000) engages in an ad hominem attack on my credibility concerning my descriptions of the course lecture material I referenced from the Backster School of Lie Detection. I do not dispute Mr. Backster's current position as it is presented in Matte (2000). However, I do maintain that my previous statements are accurate as they were presented. I attended the Backster School during the fall of 1976 as part of class PE-68. At that time, the section on question formulation was given by Charles Hess, not by Mr. Backster. Mr. Hess' lectures included a discussion of the practice of stimulating the comparison questions between repetitions. Since the lectures were not videotaped, I cannot prove that we were given those lectures. It has been my practice throughout my 24 years in the polygraph profession to review the comparison questions and relevant questions with subjects, both in the laboratory and in real cases. Over those years, I have not had any difficulty in finding subjects deceptive, when the data warrant (Honts, 1997).

However as a bottom line, neither Mr. Backster's opinions, nor my field practices matter when it comes to the scientific evaluation of the validity of the review of comparison questions between repetitions. Data, not appeals to authority, are the controlling factor in science. The data clearly support the review of comparison questions between question repetitions in both probable-lie and directed-lie comparison question tests.

Summary

All of the published data from actual polygraph examinations (not imagined ones), both from the laboratory and from the field, support the review of questions between charts, even if only the comparison questions are discussed (Dawson, 1980; Patrick & Iacono, 1989) and even with directed-lie comparison questions (Honts & Raskin, 1988; Horowitz et al., 1997). Readers who are interested in data, not unsupported speculation, are referred to Honts (1999) and to Honts el al., (2000).

References

Abrams, S. (1999). A response to Honts on the issue of the discussion of questions between charts. *Polygraph*, 28, 223-229.

Dawson, M. E. (1981). Physiological detection of deception: Measurement of responses to questions and answers during countermeasure maneuvers. *Psychophysiology*, 17, 8-17.

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

Honts, C. R. (1997, May). Is it time to reject the friendly polygraph examiner hypothesis (FEPH)? Paper presented at the annual meeting of the American Psychological Society, Washington, D.C. Also available online at: http://truth.boisestate.edu/polygraph/fpeh.html

Honts, C. R. (1999). The discussion of questions between list repetitions (charts) is associated with increased test accuracy. *Polygraph*, 28, 117-123.

Honts, C. R., & Gordon, A., (1998). A critical analysis of Matte's analysis of the directed lie. *Polygraph*, 27, 241-252.

Honts, C. R., & Raskin, D. C. (1988). A field study of the validity of the directed lie control question. *Journal of Police Science and Administration*, 16, 56-61.

Honts, C. R., Raskin, D. C. Amato, S. L., Gordon, A. & Devitt, M. (2000). The hybrid directed-lie test, the overemphasized comparison question, chimeras and other inventions: A rejoinder to Abrams (1999). *Polygraph*, 29, 156-168.

Horowitz, S. W., Kircher, J. C., Honts, C. R., & Raskin, D. C. (1997). The role of comparison questions in physiological detection of deception. *Psychophysiology*, 34, 108-115.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Iacono, W., & Lykken, D. (1997). The scientific status of research on polygraph techniques: The case against polygraph tests. Chapter in, D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.) *Modern scientific evidence: The law and science of expert testimony* (pp. 582-618).

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

Matte, J. A. (2000). A critical analysis of Honts' study: The discussion (stimulation) of comparison questions. *Polygraph*, 29, 146-149.

Matte, J. A., & Reuss, R. M. (1999). Validation of potential response elements in the directed-lie control question. *Polygraph*, 28, 124-142.

Office of Technology Assessment (1983). Scientific validity of polygraph testing: A research review and evaluation. OTA-TM-H-15. Washington, DC: U. S. Congress, Office of Technology Assessment.

Patrick, C. J. & Iacono W. G. (1989). Psychopathy, threat, and polygraph test accuracy. *Journal of Applied Psychology*, 74, 347-355.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (1997). The scientific status of research on polygraph techniques: The case for polygraph tests. Chapter in, D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.) *Modern scientific evidence: The law and science of expert testimony* (pp. 565-582).

Szucko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist*, 36, 488-496.

Polygraph and Investigation, Perfect Partners: A Case Study

Robert J. Drdak

Key words: case study

The exclusive mountain resort located 30 miles west of Boone, North Carolina resembles a picture postcard. Nestled in the shadow of Grandfather Mountain, its perfectly manicured golf course is surrounded by custom homes, tennis courts, and mountain shrubbery. Many of the homes, some costing over one million dollars, are weekend or vacation homes and sit vacant for much of the year. The community is gated and security is tight. A guardhouse controls access and armed guards patrol the property twenty-four hours a day. Most houses have electronic security systems installed. To facilitate security, residence keys and burglar alarm codes are maintained by the security force at the guardhouse.

In March 1996, one house in this custom neighborhood was set apart from the others. Perched on a hill, long and low with a shake roof, it was a perfect example of mountain architecture. But what made the house unique was the massive mahogany-trimmed trophy room filled with treasures from a lifetime of big game hunting around the world. Twenty-five foot ceilings and a walk-in stone fireplace also made the room a perfect showcase for the owner's world-class collection of modern and antique weapons valued at over one million dollars. Many of the firearms were one of a kind pieces owned by historical figures such as Napoleon, Wild Bill Hickok, and the Guinness Brothers.

On Saturday morning, March 23, 1996, the owner's maid of ten years entered the home to do a routine check and cleaning. She found everything in order and left after turning off all the lights and activating the alarm.

During the evening and midnight shift, the patrol guard noticed that an outside and some inside lights were on at the home. Observing nothing suspicious, the guards assumed the owners were in residence, and made no attempt to check further. This was the standard procedure for the guard force. Also on this particular night, the midnight patrol guard noticed that a gate leading from the resort maintenance facility to the residential area was unlocked. Other than the guarded front gate, this was the only other entrance to the residential area of the resort. Assuming the gate was inadvertently left open by a maintenance employee, the guard closed and locked it.

The lights were observed to be on at the home through Tuesday night. On Wednesday morning, the chief of security went to the house for a routine check and found a windowpane broken on a rear door. The door was unlocked and the alarm was deactivated. With great apprehension, he entered the home and found the owner's entire gun collection missing.

Local authorities were immediately notified. Due to the size of the theft and the close proximity to the Tennessee state line, it was assumed that some or all of the missing guns would be transported interstate. With this in mind, the local authorities contacted the Charlotte Field Office of the Federal Bureau of Investigation (FBI) and requested assistance. An aggressive Major Theft Investigation was initiated and the FBI Evidence Response Team along with a squad of Special Agents was dispatched to the resort where a command post was established.

The author retired from the Federal Bureau of Investigation in December 1999 after 28 years service. He currently is President and Director of DrD Consulting Group, a professional company specializing in polygraph testing and related matters. Reprint requests should be sent to P.O. Box 7366, Charlotte, N.C. 28241 or e-mailed to rdrdak@yahoo.com.

The FBI considers the polygraph to be a valuable investigative tool that, when used in conjunction with traditional investigative techniques, can quickly and economically focus an investigation by eliminating innocent suspects and identifying the guilty. This is exactly what the case agent in this investigation had in mind when he called me at home on Wednesday night and requested that I come to the resort as soon as possible. I arrived at the command post on Thursday morning and was quickly integrated into the investigation.

As information was gathered and analyzed, a picture of the crime soon developed. It appeared that one or more people used a key to enter the home and then used the code to disarm the alarm. The thief, or thieves, targeted only the gun collection, leaving over $100,000 worth of silver flatware in an unlocked safe. The thief or thieves also cut the telephone lines to the house and somehow managed to enter and leave the resort undetected.

Based on the initial investigation, it was apparent that the subject(s) in this theft had knowledge of the gun collection, had access to the house key and alarm codes, and had a way to enter and leave the resort grounds without suspicion. It appeared the theft actually occurred during the Saturday night-Sunday morning period. Based on this, a list of possible suspects was formulated. A decision was made to polygraph the two guards who were on duty during the Saturday-Sunday midnight shift. The examinations were scheduled for Friday morning.

Thursday's investigation yielded another suspect. Larry Neal Lane was a former security guard at the resort that left under unfavorable circumstances and was known to harbor resentment toward the resort. He had previous brief employment in law enforcement and now was a licensed private investigator in North Carolina. Lane was also a gun enthusiast who frequented area gun shows and bought and sold low and medium quality firearms. While working as a security guard at the resort, he had access to alarm codes and the keys to homes at the resort.

Lane made an uncharacteristic visit to the resort security guard house the day before the gun theft was discovered. He was cordial and in good spirits and said he stopped by to "see what was going on". Those present thought the visit to be strange, considering his feelings toward the resort management.

Agents immediately began to gather background information on Lane. He was contacted and agreed to come to the command post on Friday afternoon to be interviewed. While Lane was being questioned at the command post, Agents were interviewing his wife concerning his activities during the past week. Agents were also dispatched to nearby Tennessee to interview a man who had observed Lane at a gun show on Sunday attempting to sell a unique rifle that was obviously out of his normal price range.

The Polygraph Examinations

On Friday morning, both security guards appeared on schedule for their polygraph examinations. Each denied any involvement in the theft. The examination format was a basic Zone Comparison utilized by the FBI. It included three comparison questions and three relevant questions formatted as follows:

1. Irrelevant Question
2. Sacrifice Relevant Question
3. Symptomatic Question
4. Comparison Question
5. Relevant Question: Did you take any of those missing guns from the (owner's name) house?
6. Comparison Question
7. Relevant Question: Did you help anyone take those guns from the (owner's name) house?
8. Comparison Question
9. Relevant Question: Do you know where any of (owner's name) guns are now?
10. Irrelevant

Both guards passed the examination with an examiner opinion of No Deception Indicated (see charts 1 - 6).

Lane arrived at the command post Friday afternoon for his interview. At the same time, unknown to him, his wife was being interviewed at another location. She was certain in her own mind that Lane was not involved in the theft and cooperated fully. She gave an honest account of Lane's activities during the critical time period, and revealed that he was "out working" until early Sunday morning.

Lane, on the other hand, was not so honest. He told the agents that he was home in bed, and insisted that he was not involved in the theft. When asked if he would take a polygraph examination, Lane seemed hesitant and stated that he did not believe in the reliability of polygraph, but agreed to meet with the examiner.

Located in the next room, I was immediately available to talk to Lane. After some discussion, he agreed to an examination. During the pretest interview, I could sense Lane's deep resentment toward the resort management. He was highly critical of the security at the resort, believing it to be mismanaged and inadequate. He stated that it was just a matter of time before a major theft like this one took place and that he hoped such a high profile case would prove his criticisms were correct. He continued to deny any involvement in the theft.

Lane was given a polygraph examination formatted exactly as the one taken by the two security guards and containing the same three relevant questions. He showed dramatic responses to the relevant questions (see charts 7 - 9). I was certain we had found our thief.

Before beginning the interrogation, I consulted with the agent who was coordinating the investigation. He informed me that Lane's wife's account of his activities on the night of the theft differed from his own. He also told me that the gun dealer interviewed in Tennessee positively identified Lane as the man who tried to sell a very expensive high-powered rifle with a unique zebra wood stock. His description of the rifle perfectly matched one of the missing guns.

The Interrogation

I am a firm believer in the use of techniques taught in most interview and interrogation classes: direct confrontation, rejection of denials, and theme development. However, nothing can replace good old evidence as ammunition in an interrogation. Most interrogations don't follow the Hollywood script seen in so many police movies and television shows where the guilty subject tells all after being glared at by two intimidating detectives. In any interrogation, the investigator tries to convince a guilty subject to make admissions against his own best interests. These admissions could result in embarrassment, lost employment, and even imprisonment. Most guilty subjects will not make such admissions unless the interrogator can convince them that the benefit of making the admissions outweighs the benefit of denials. This is when the collateral investigation can make the polygraph examiner look good.

Minimize, rationalize, and project blame - three of the basic ingredients in a successful interrogation - screamed to be used in the present situation. Lane had already projected the blame for the theft onto the security administration of the resort. By convincing him that he had done a noble thing in exposing the security problems at the resort, and minimizing the consequences provided the guns were returned, I was able to slowly break through Lane's denials and get small hints of his guilt. I described to him, based on our investigation, how the crime took place. When I finally confronted him with the results of the interviews of his wife and the witness to his attempted gun sale in Tennessee, his denials became weaker and weaker.

After several hours, Lane gave a complete confession and executed a consent to search his home and automobiles. He led us to all the guns, which were hidden in a derelict car parked near his house. During the search we observed other items which had apparently been stolen from other residences in the resort. Local authorities were notified and a subsequent search resulted in local as well as federal charges being filed against Lane.

During the interrogation after Lane admitted his guilt, he was asked to provide a detailed account of how he performed the crime. He stated, "Why? You already know everything". This was an offhanded tribute to the value of our initial investigation. It was so thorough that Lane was convinced further denials would be fruitless. I feel this played a major role in our success, and demonstrates how polygraph coupled with investigation can be a formidable tool.

Lane did provide us with the details of the theft. He stated that while he was working at the resort, he had admired the valuable gun collection. Before leaving his job there, he made a copy of the keys to the house and noted the alarm code. He also made a copy of the key to the lock on the maintenance area gate. On the night of the theft, he entered the resort through the maintenance area and left the gate unlocked to facilitate his exit. At the house, he used the copied key to enter and then disarmed the alarm using the code. He was afraid the alarm code might have changed, so he cut the phone wires as a precaution. He intentionally turned on the lights to make it appear the residents were home. Knowing the guards' patrol procedures, he timed his exit to avoid detection.

Despite Lane's planning and care, the burglary did not go without a hitch. While loading the guns into his car, the back door to the house accidentally closed, locking his keys inside. Lane had to break a windowpane so that he could re-enter the house. It was this broken pane that was noticed by the security guards that triggered the investigation.

Lane was never convicted for his crime. He took his own life days before beginning his trial on state charges.

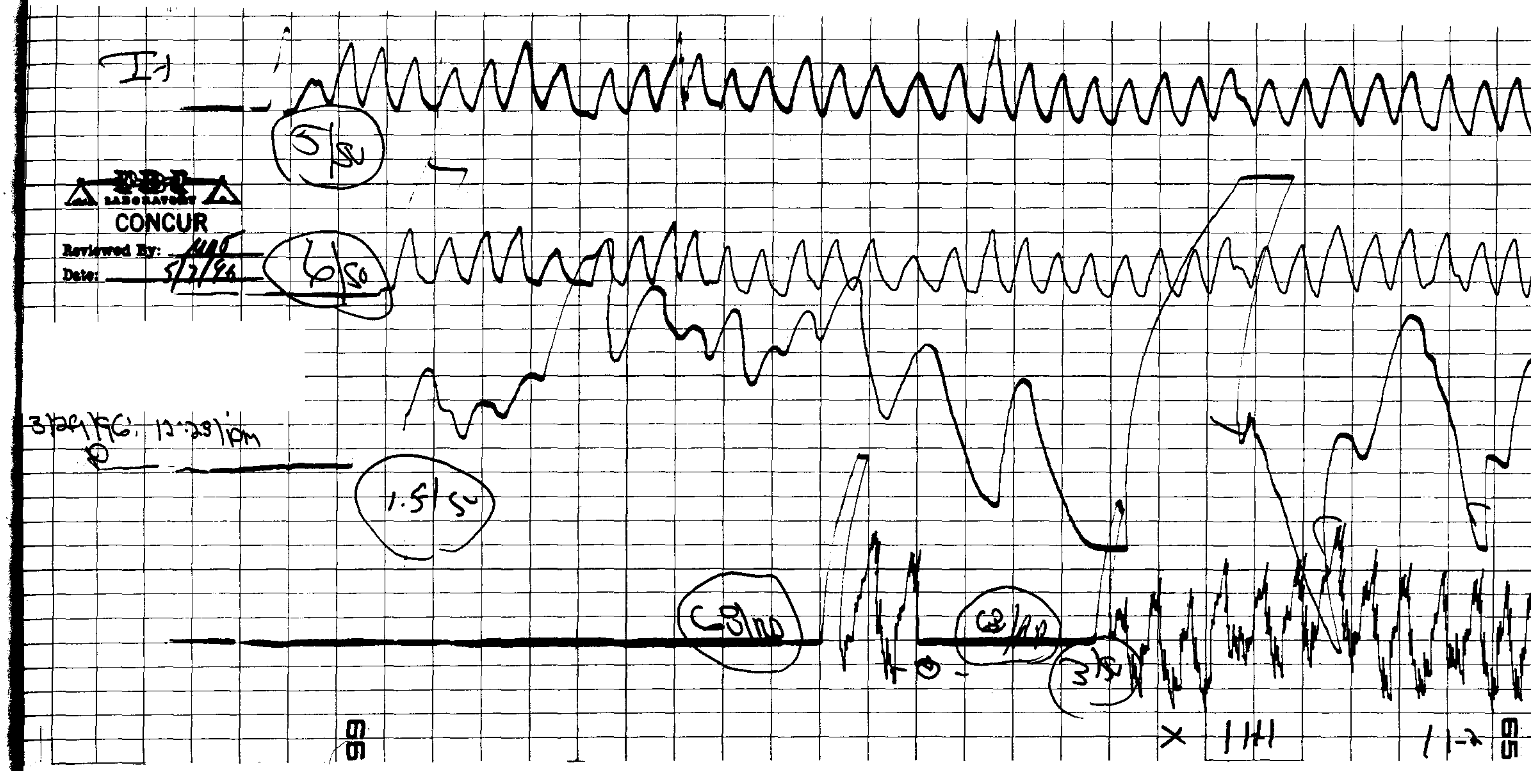
* * * * * *

Erratum

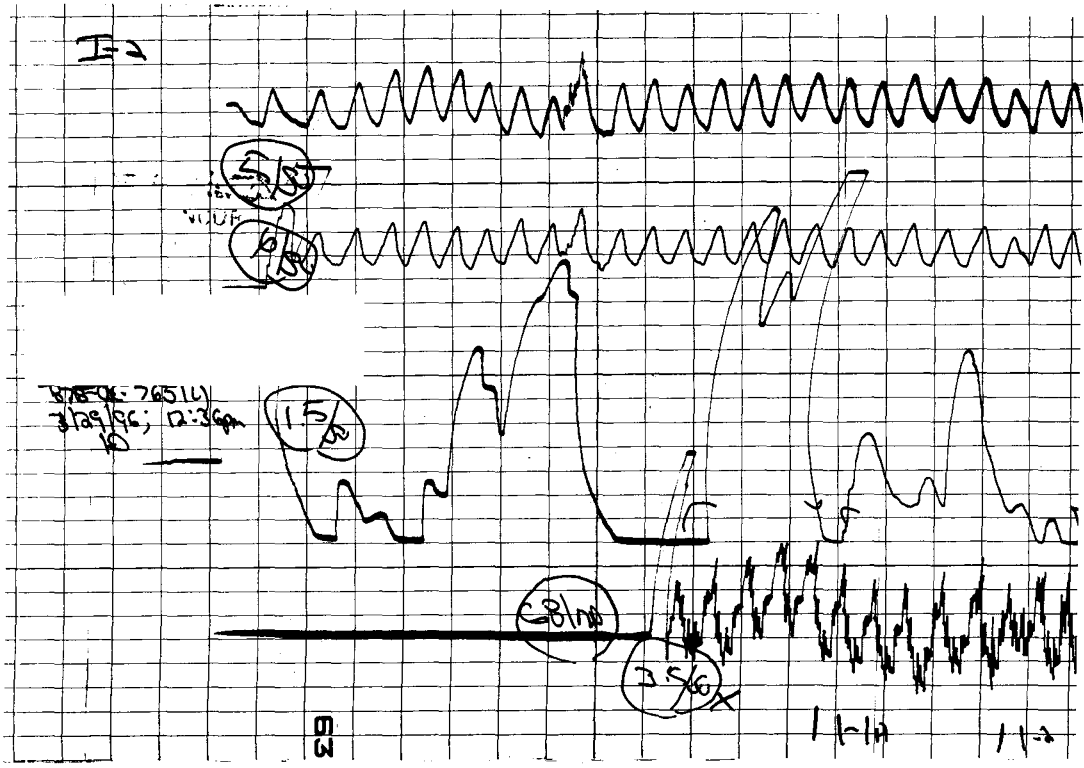
In the last edition of *Polygraph*, Table 3 (page 240) of the Dollins, Krapohl and Dutton article was formatted incorrectly. The correct information is provided below. We regret the error.

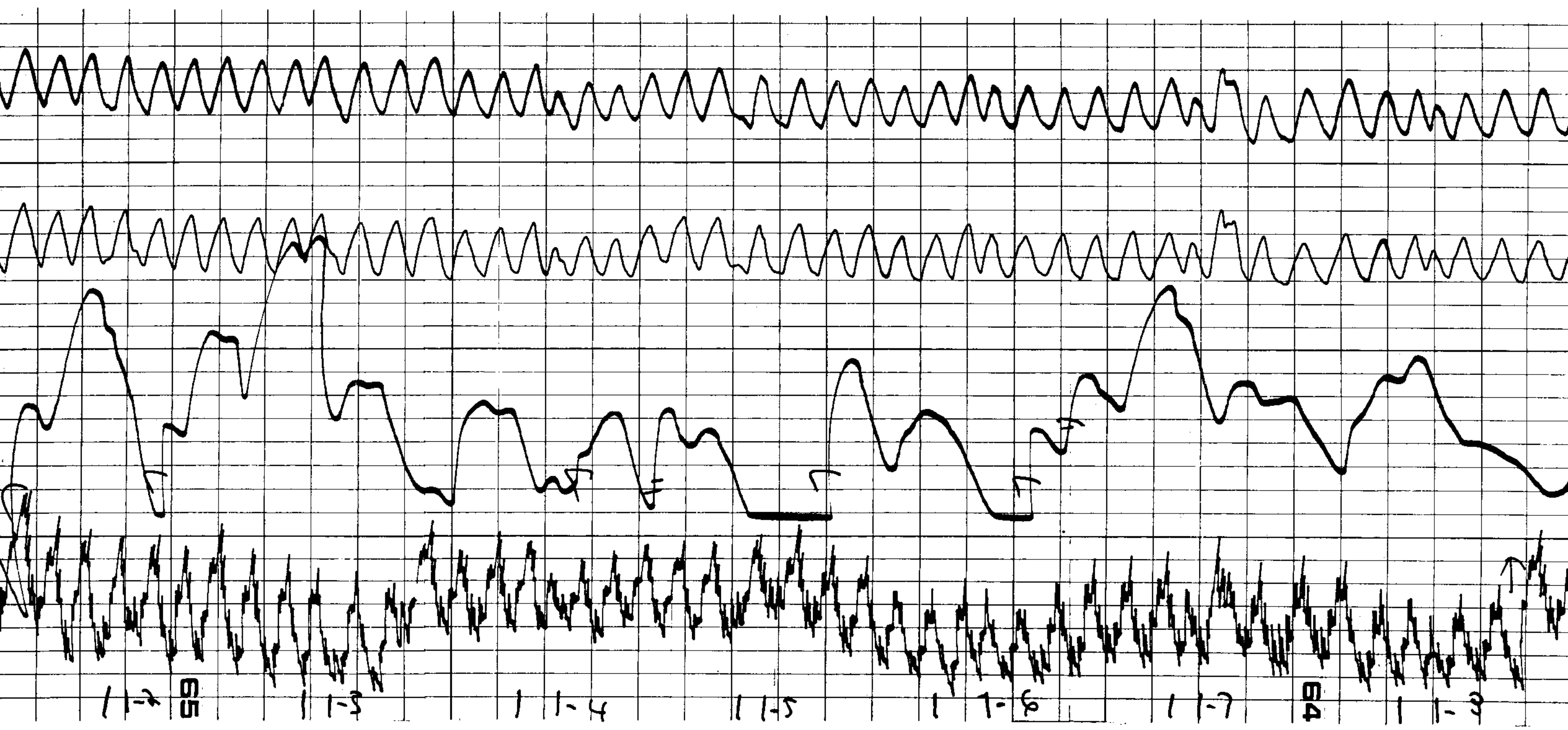
Table 3
Proportion of Agreement Between Pairs of Scoring Systems (n = 97)

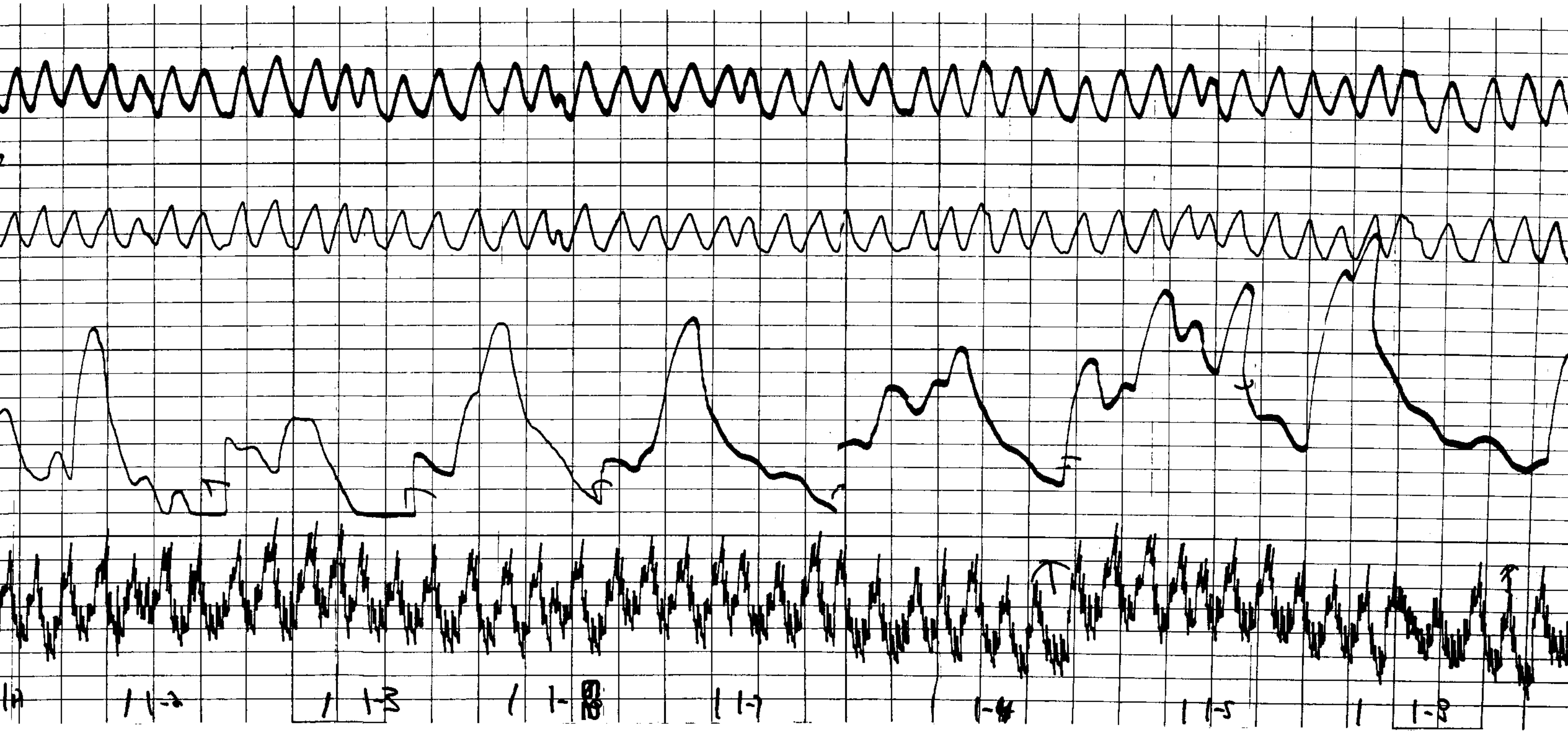
Computer Program	Chart Analysis	CPS	Identifi	PolyScore
AXCON	.907	.753	.804	.856
Chart Analysis		.742	.784	.804
CPS			.722	.753
Identifi				.722

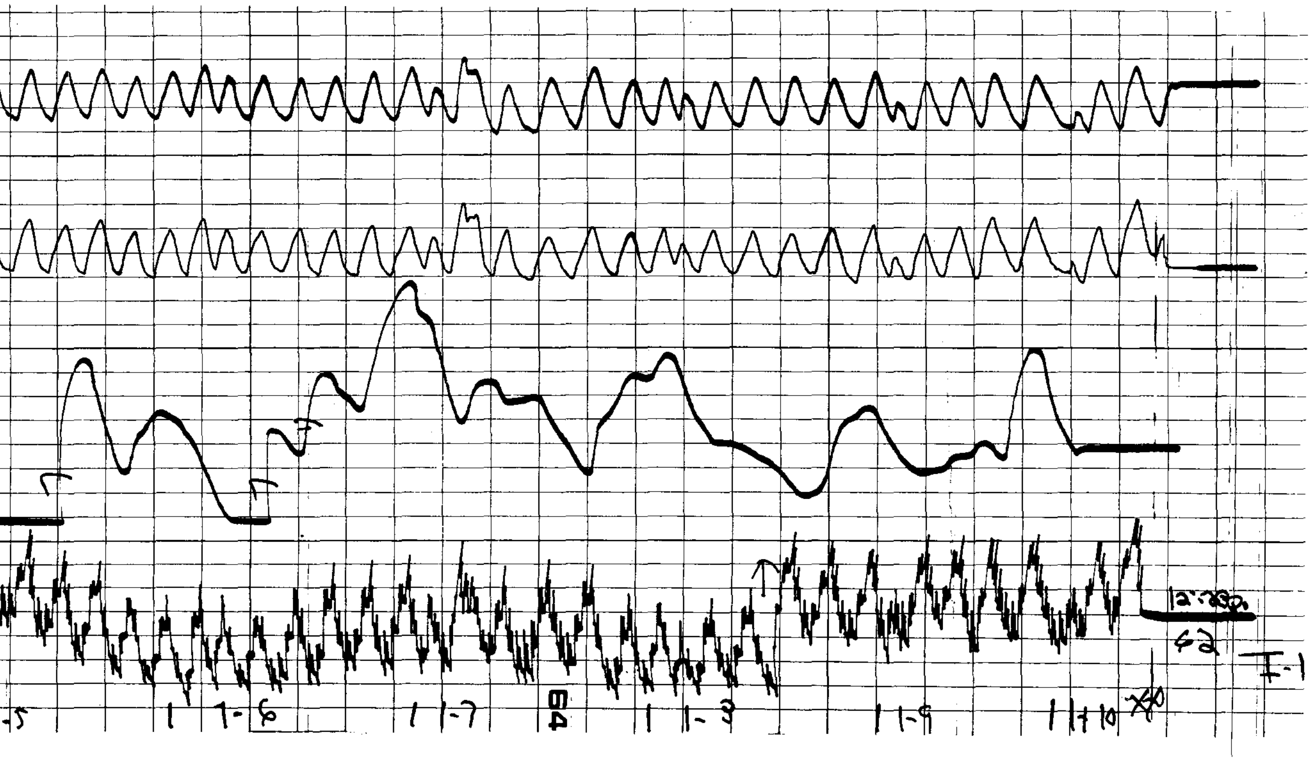


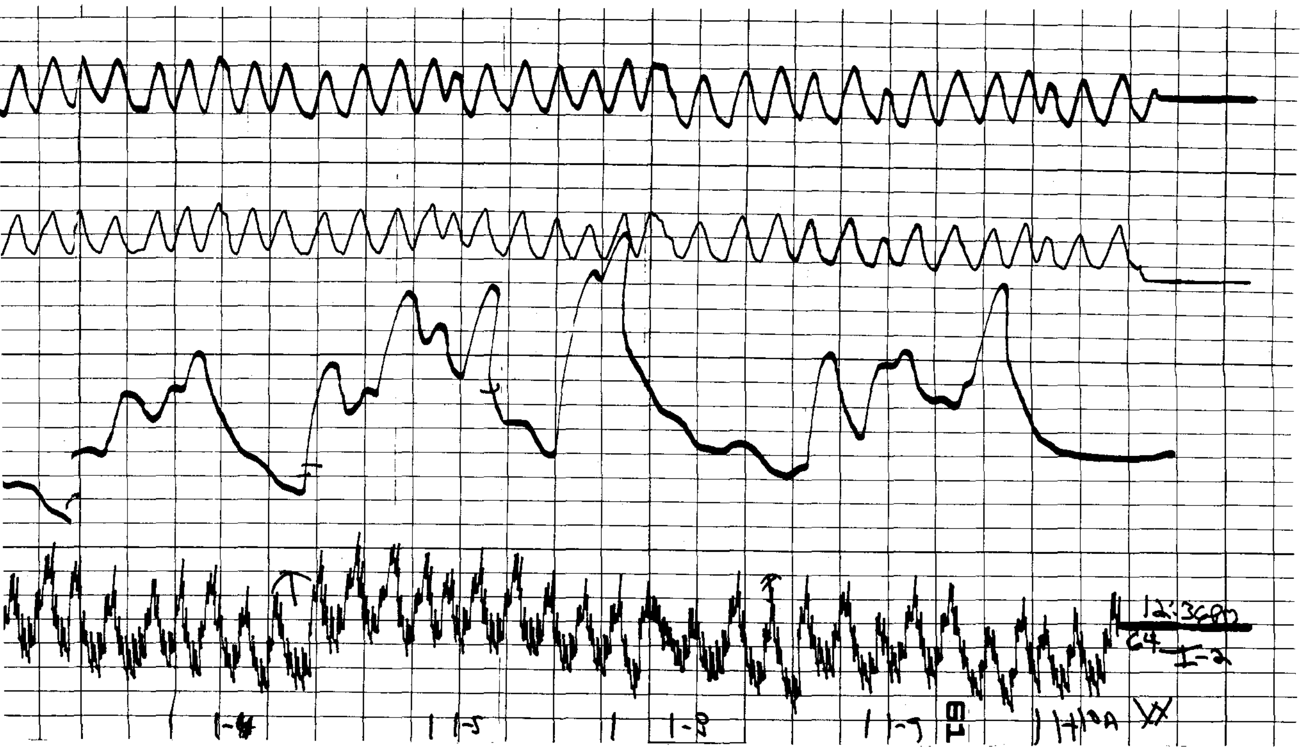
FBI LABORATORY
CONCUR
Reviewed By:
Date: 5/7/96

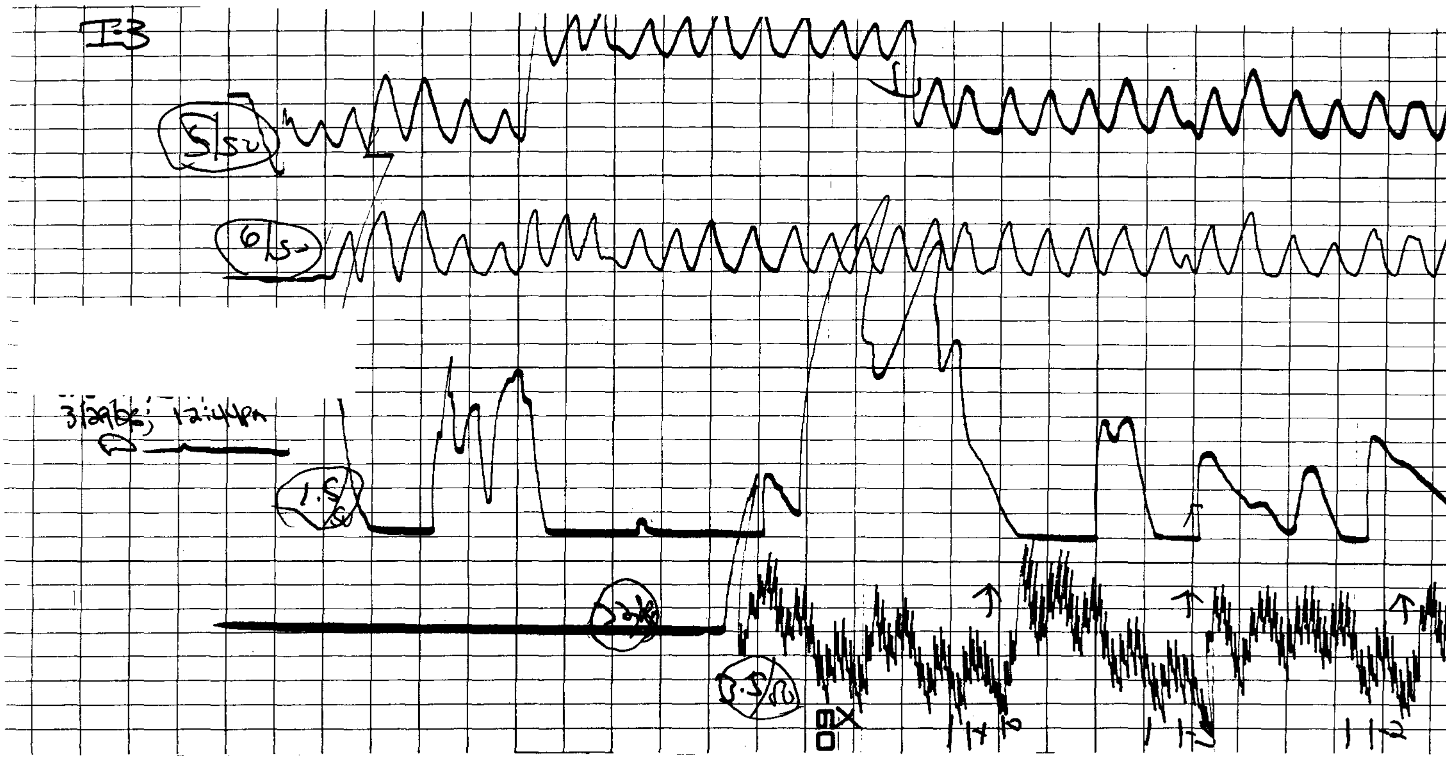


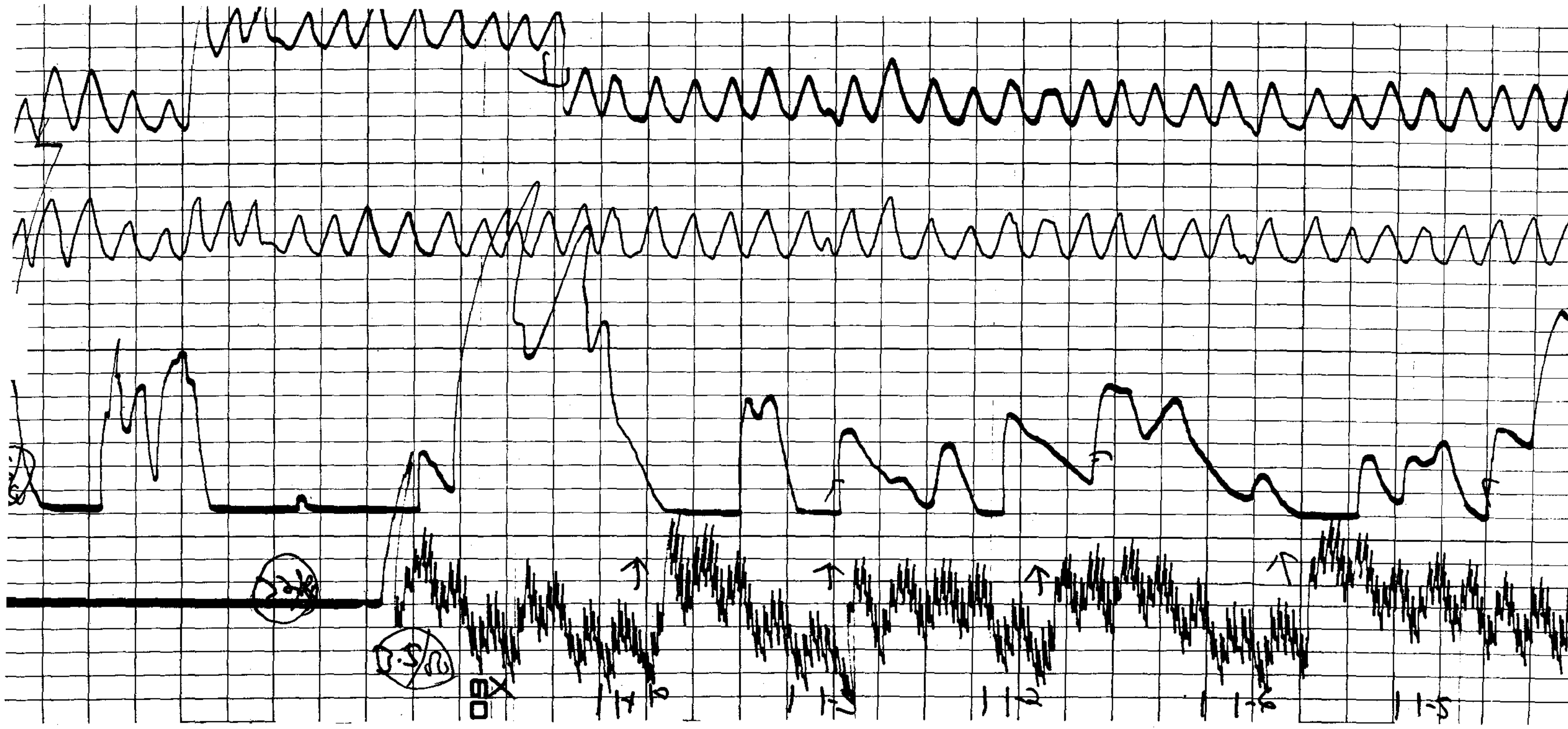

1 1-2
65
1 1-3
1 1-4
1 1-5
1 1-6
1 1-7
64
1 1-3

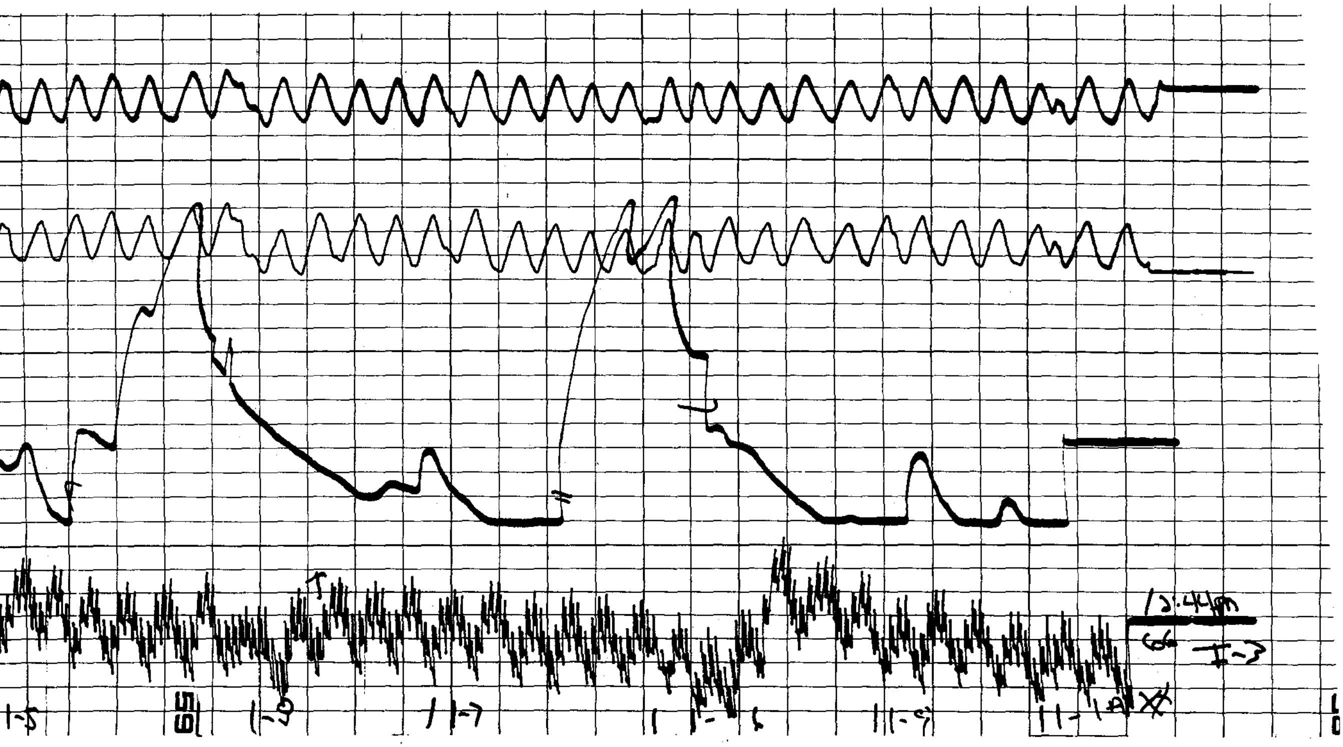

1-1
1 1-2
1 1-3
1 1-8
1 1-7
1 1-4
1 1-5
1 1-8

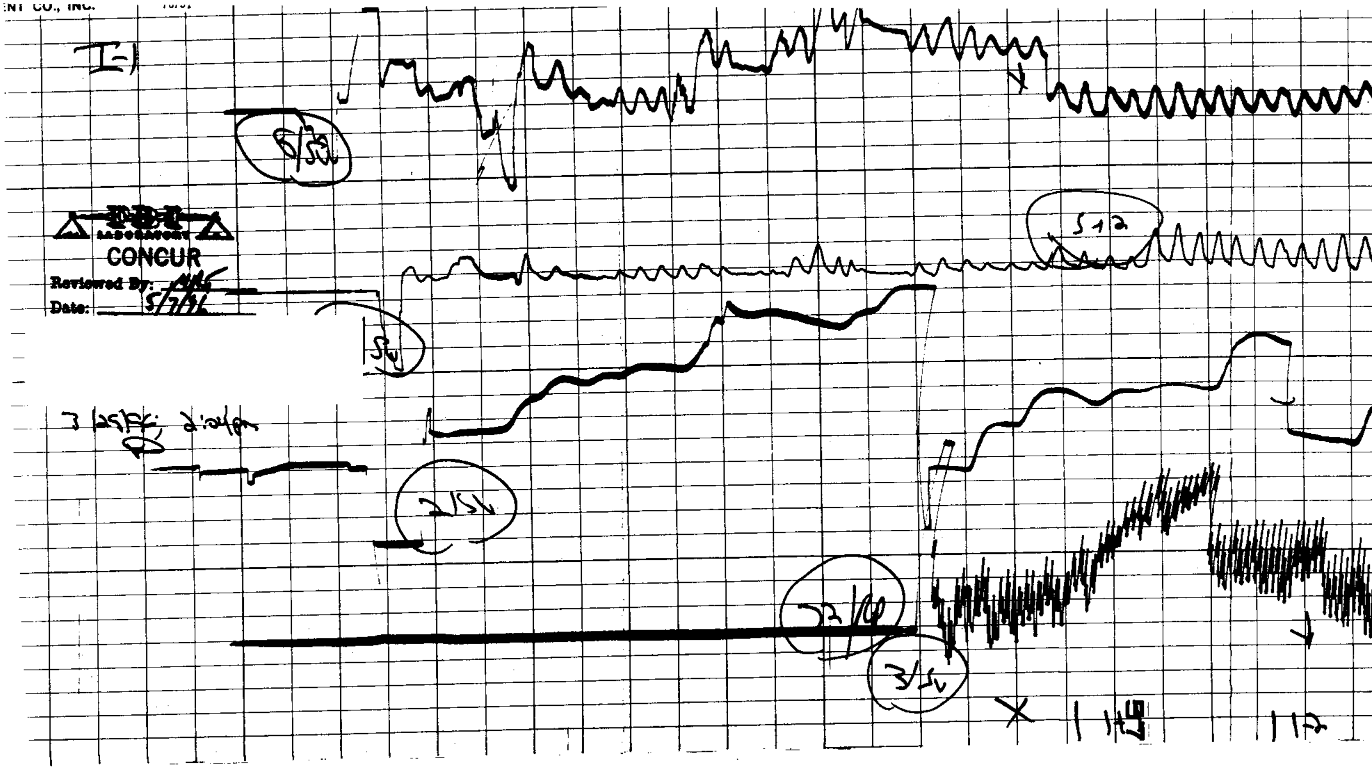

-5
1-6
1-7
E4
1-8
1-9
1+10
X
12:28p
62
I-1


1-4
1-5
1-8
1-7
E1
1+10A
XX
12:36pm
64
I-2

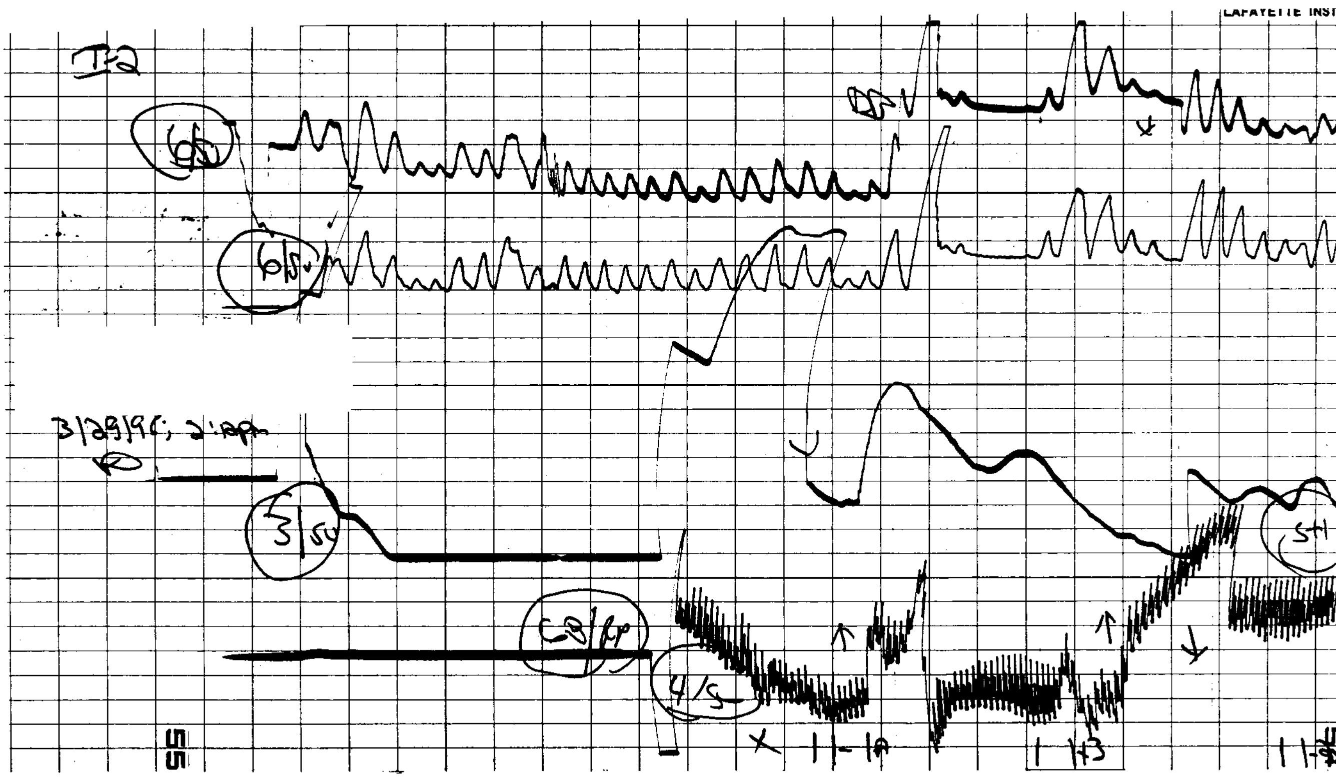

I-3
5|su
6|su
3/29/06; 12:44pm
1.5
0.5
80X
14 10
1 1-
1 1-2

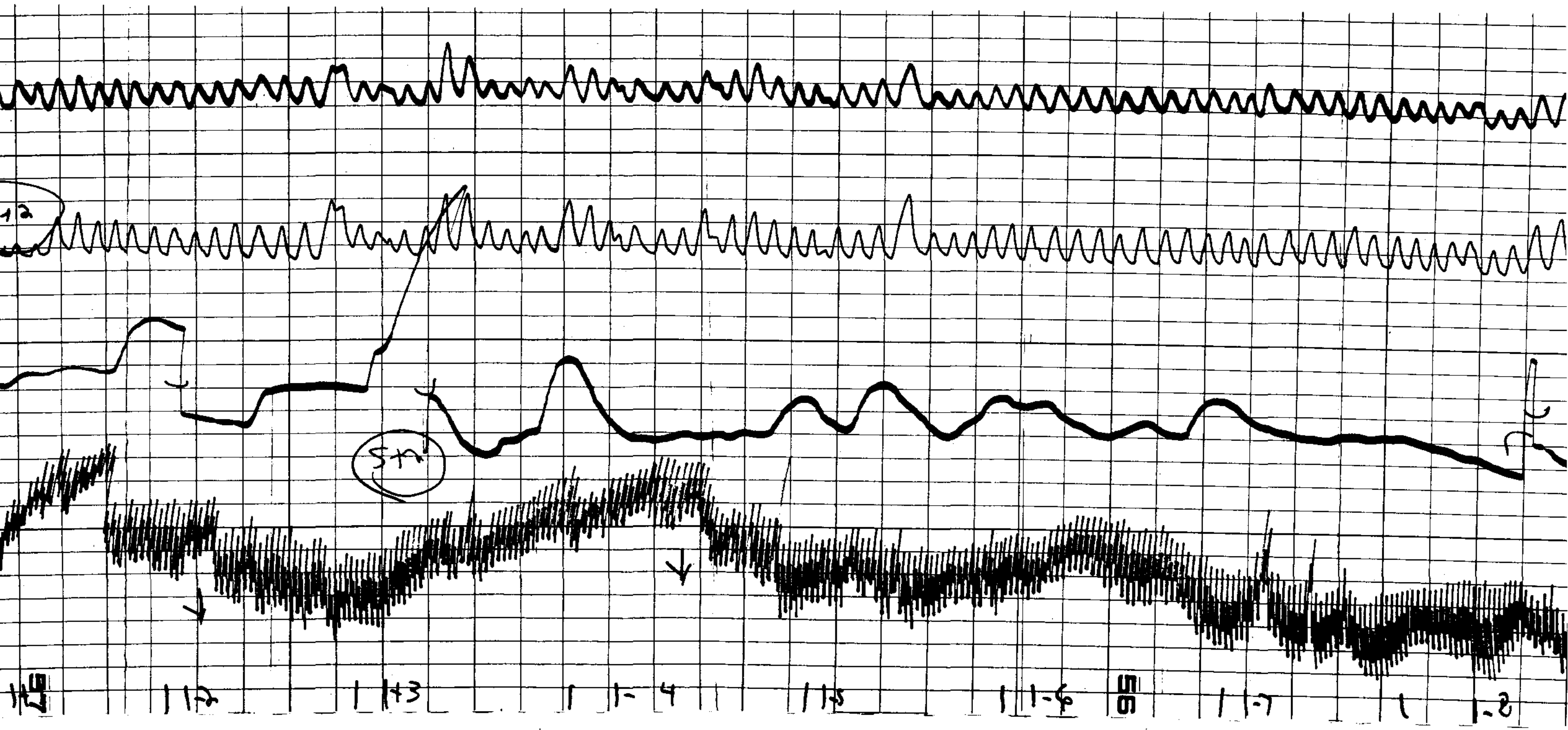

X 80
1 1-6
1 1-2
1 1-6
1 1-5


1-5
65
1-9
11-7
11-8
11-9
11-10
12:44pm
66
I-3

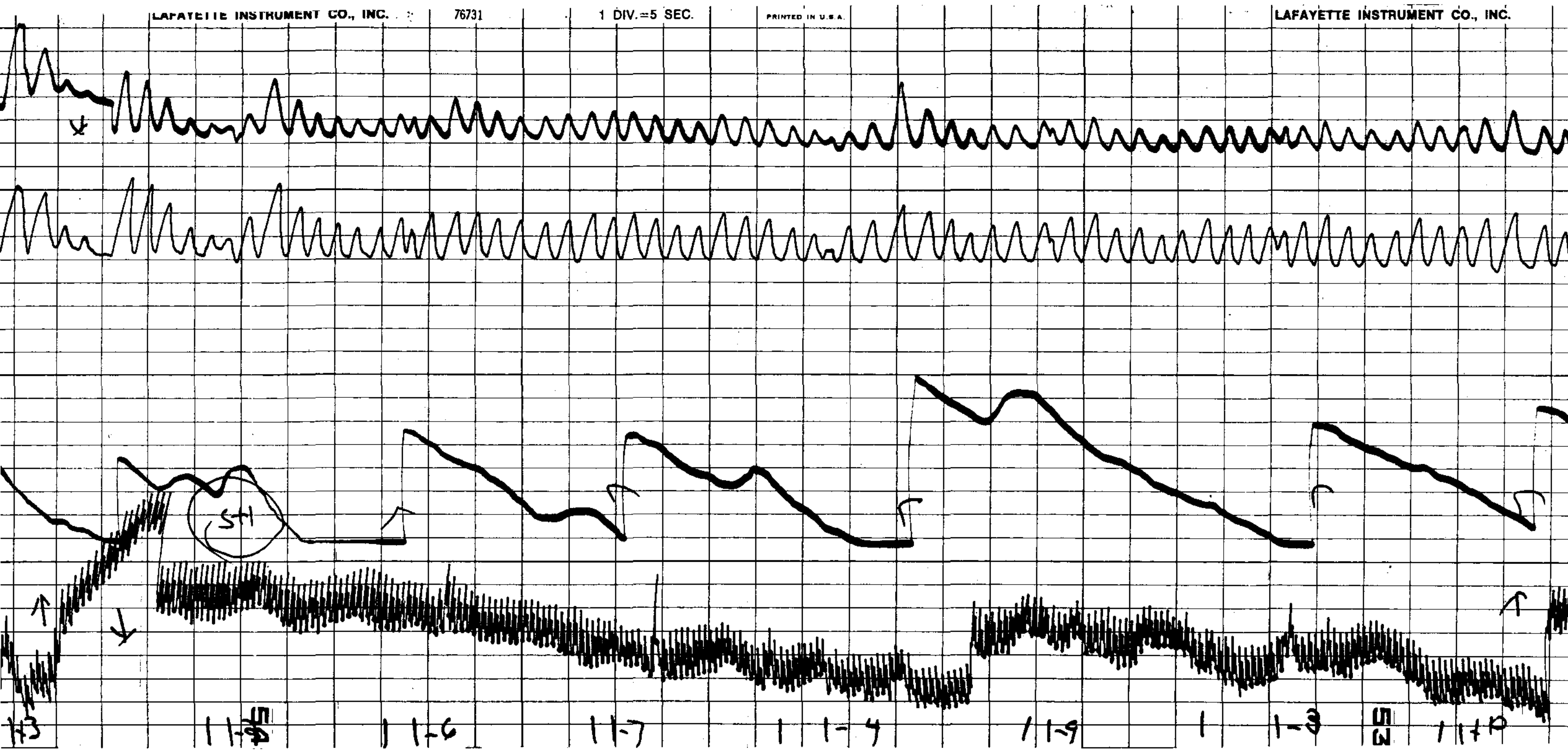


FBI LABORATORY
CONCUR
Reviewed By: MAC
Date: 5/7/96

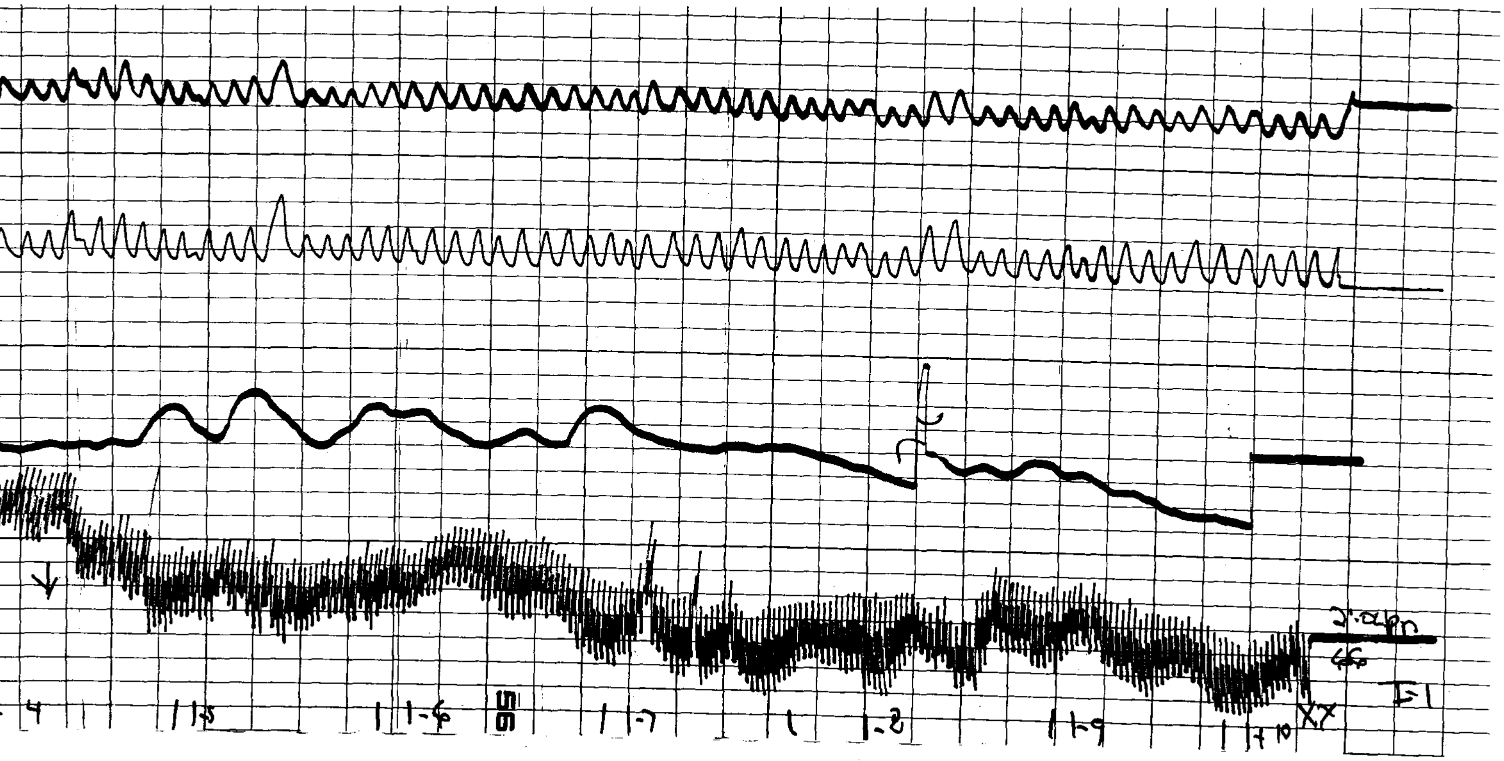


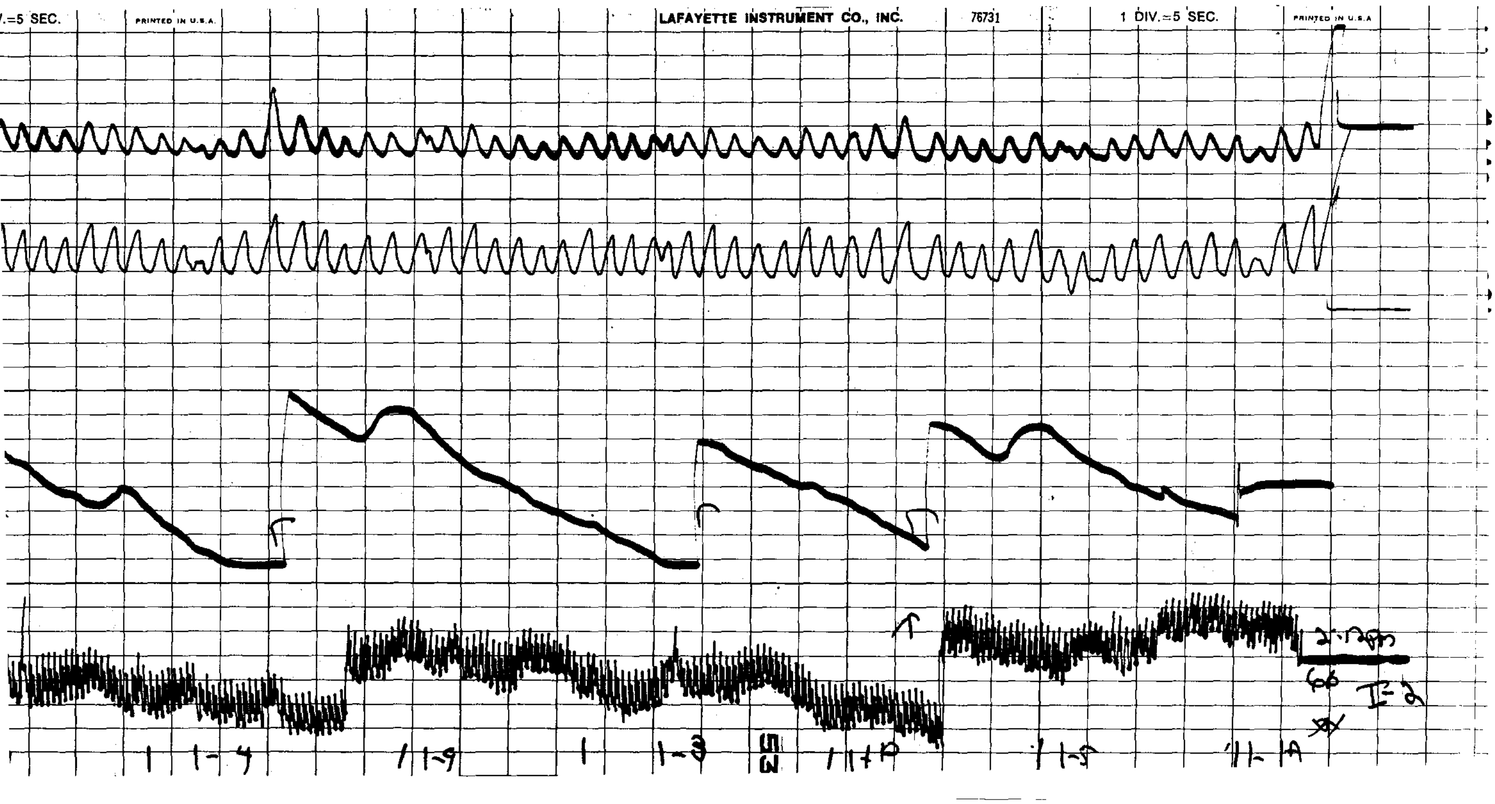


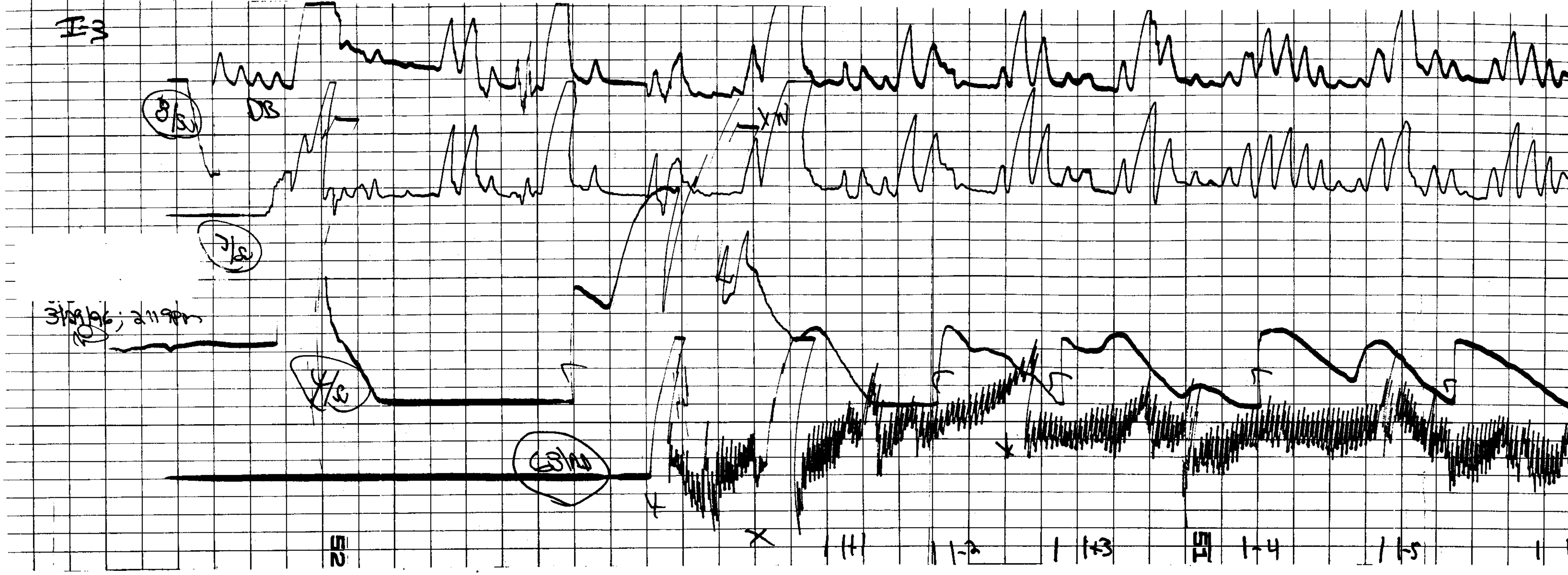
12
St
1-2
1+3
1-4
1-5
1-6
56
1-7
1-8

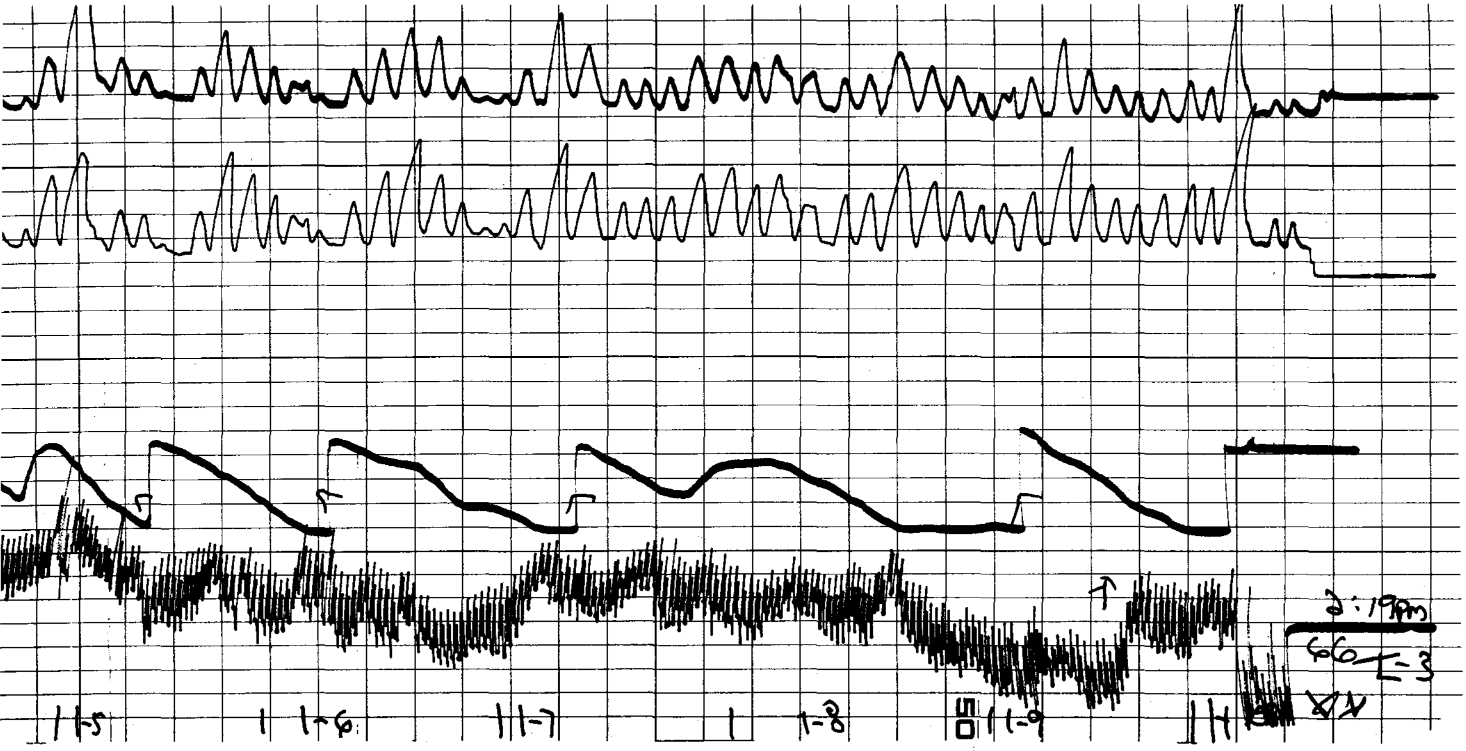


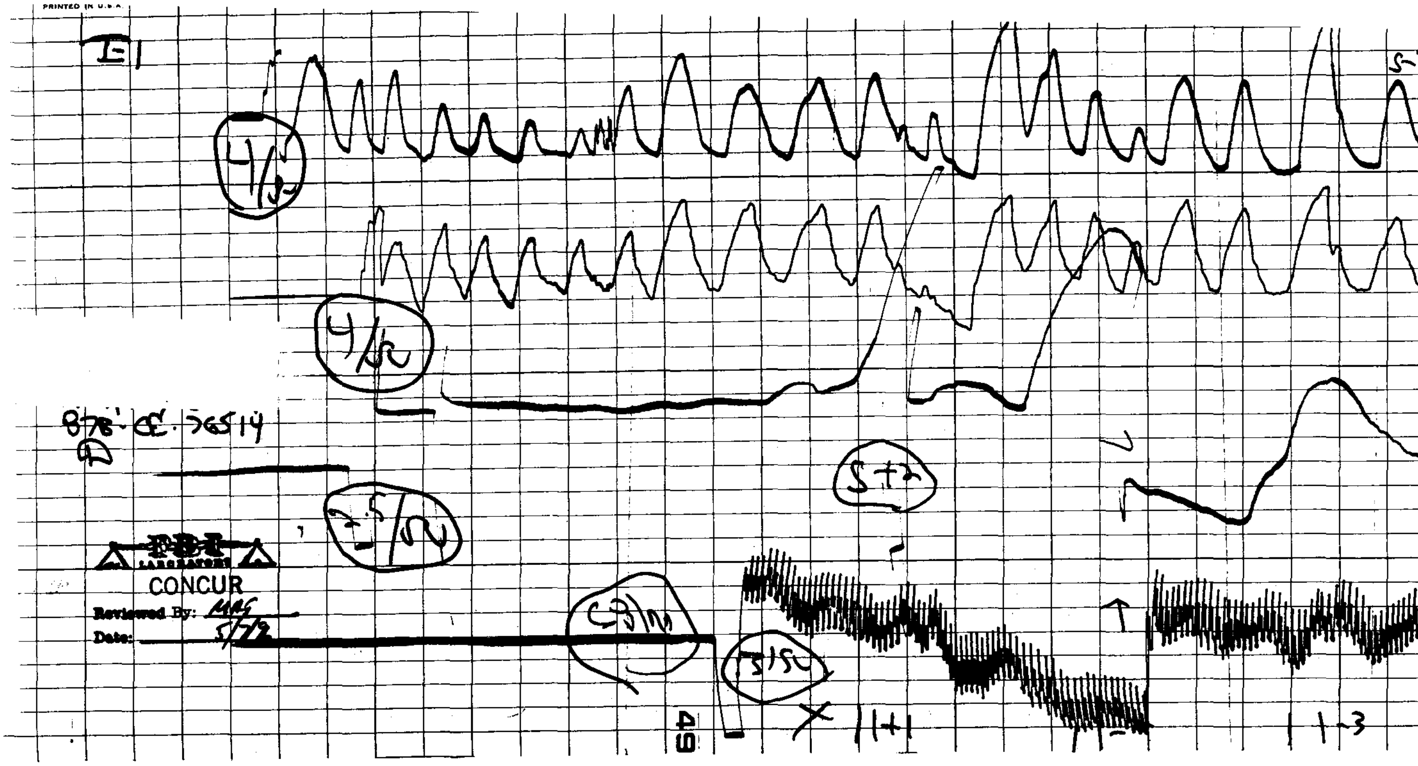
LAFAYETTE INSTRUMENT CO., INC.
76731
1 DIV.=5 SEC.
PRINTED IN U.S.A.
LAFAYETTE INSTRUMENT CO., INC.
St
1-3
1-5
1-6
1-7
1-4
1-9
1-3
53
1-10


4
11-5
11-6
56
11-7
1
1-8
11-9
11+10
XX
2:20pm
66
I-1

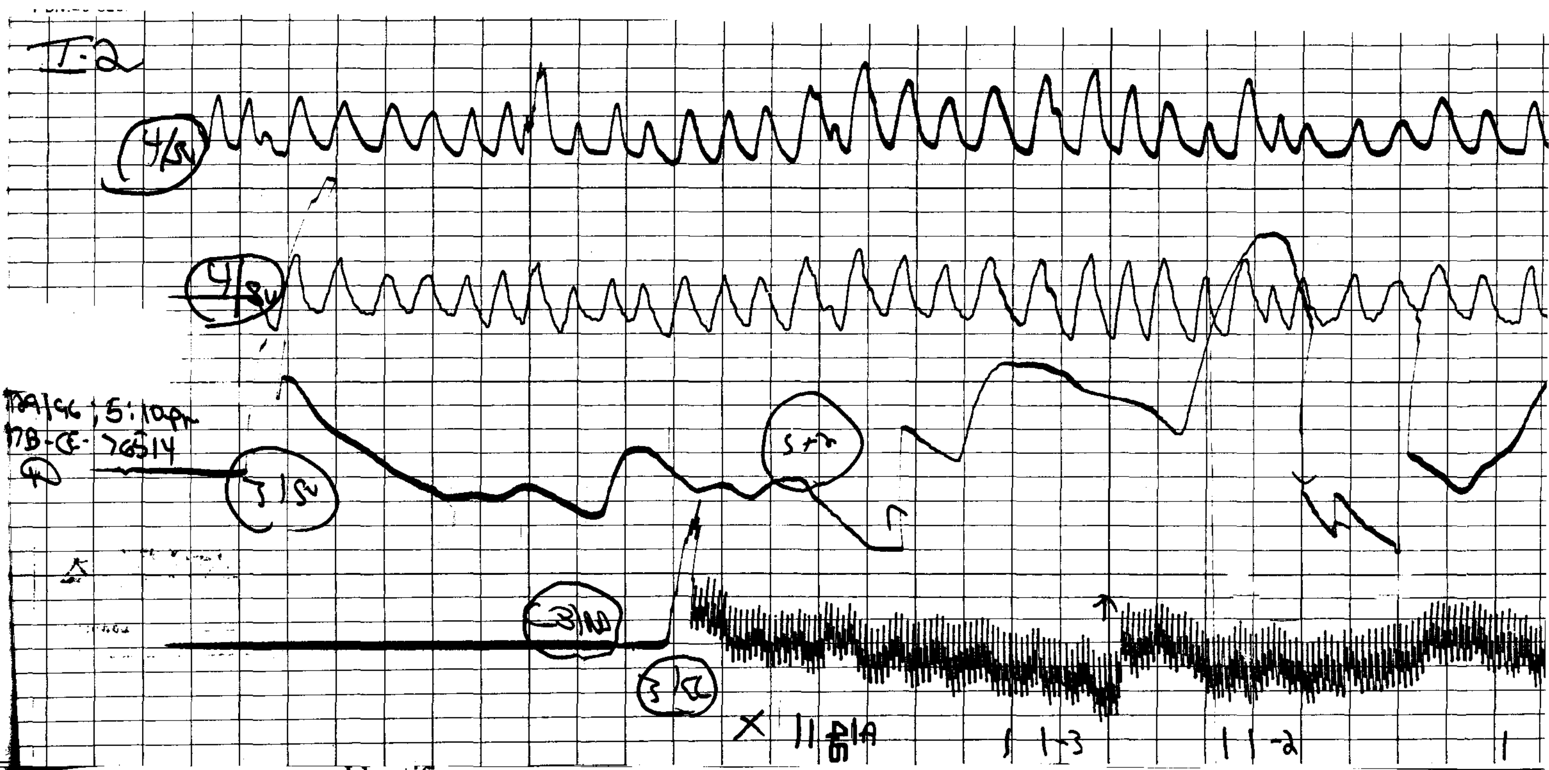

/.=5 SEC.
PRINTED IN U.S.A.
LAFAYETTE INSTRUMENT CO., INC.
76731
1 DIV.=5 SEC.
PRINTED IN U.S.A.
1 1-4
11-9
1
1-3
53
11+P
11-5
11-1A
2:12pm
66
I-2
XX

I-3
8/32
DB
7/32
XN
3/28/96; 2:19pm
4/32
68/14
52
X
+
X
1-1
1-2
X
1-3
51
1-4
1-5

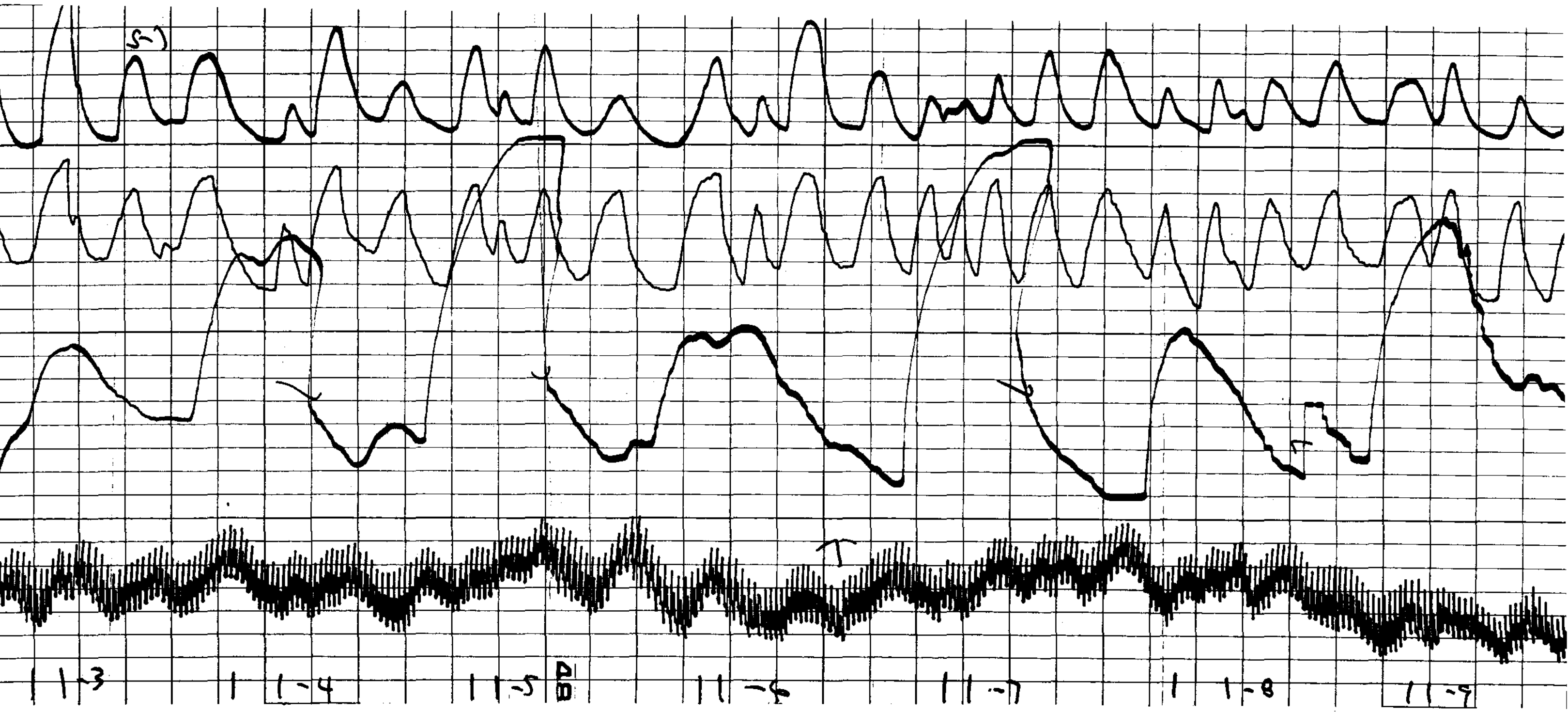

11-5
1-6
11-7
1-8
50
11-9
2:19pm
66
E-3

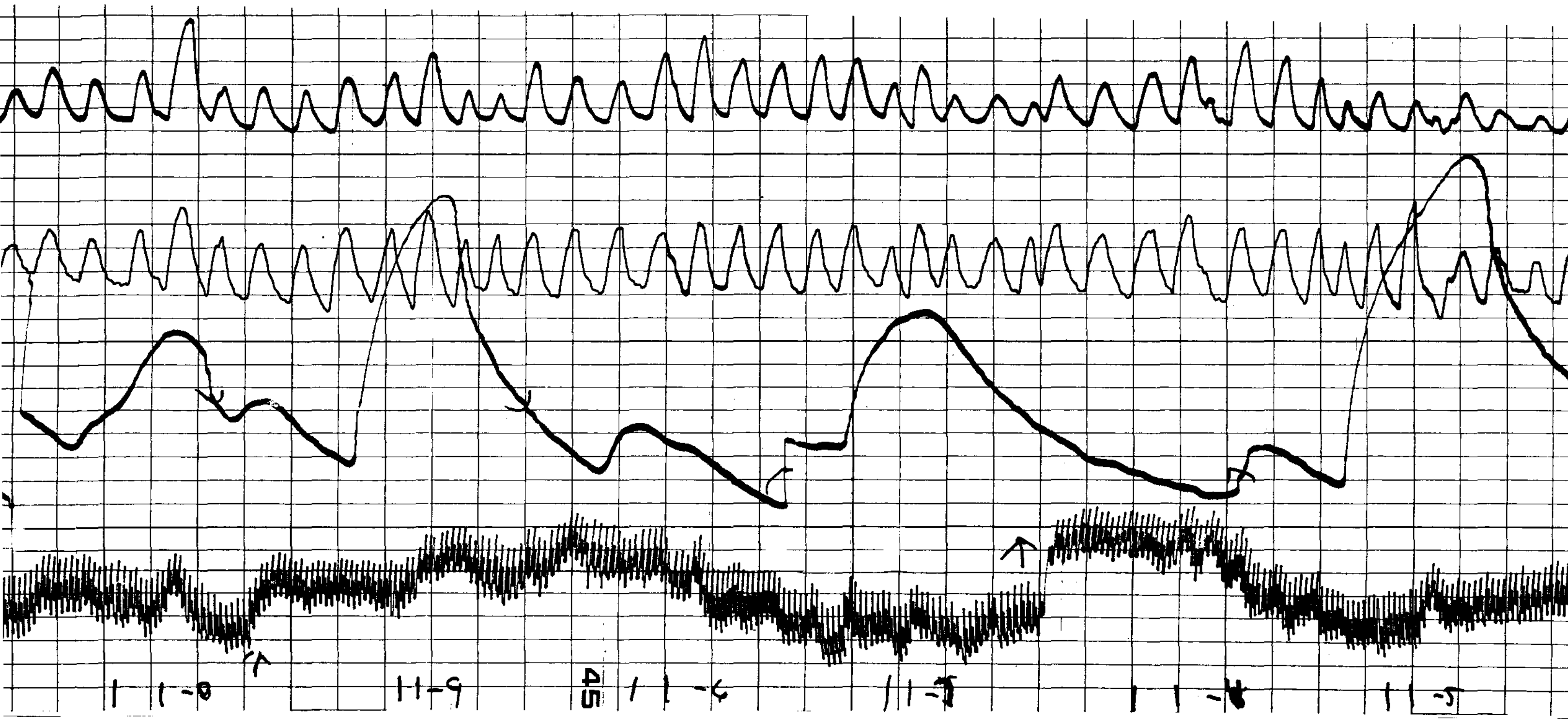


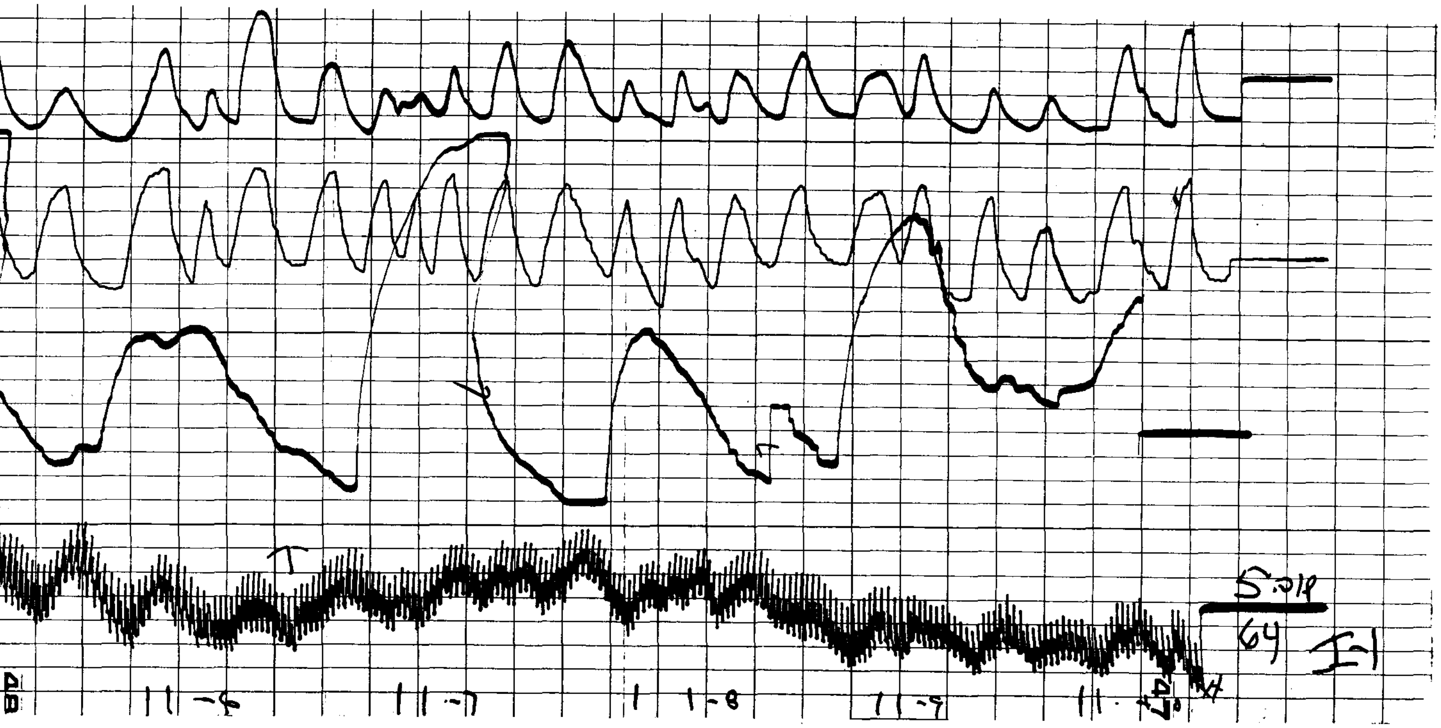
PDI LABORATORY
CONCUR
Reviewed By: MAG
Date: 5/7/2

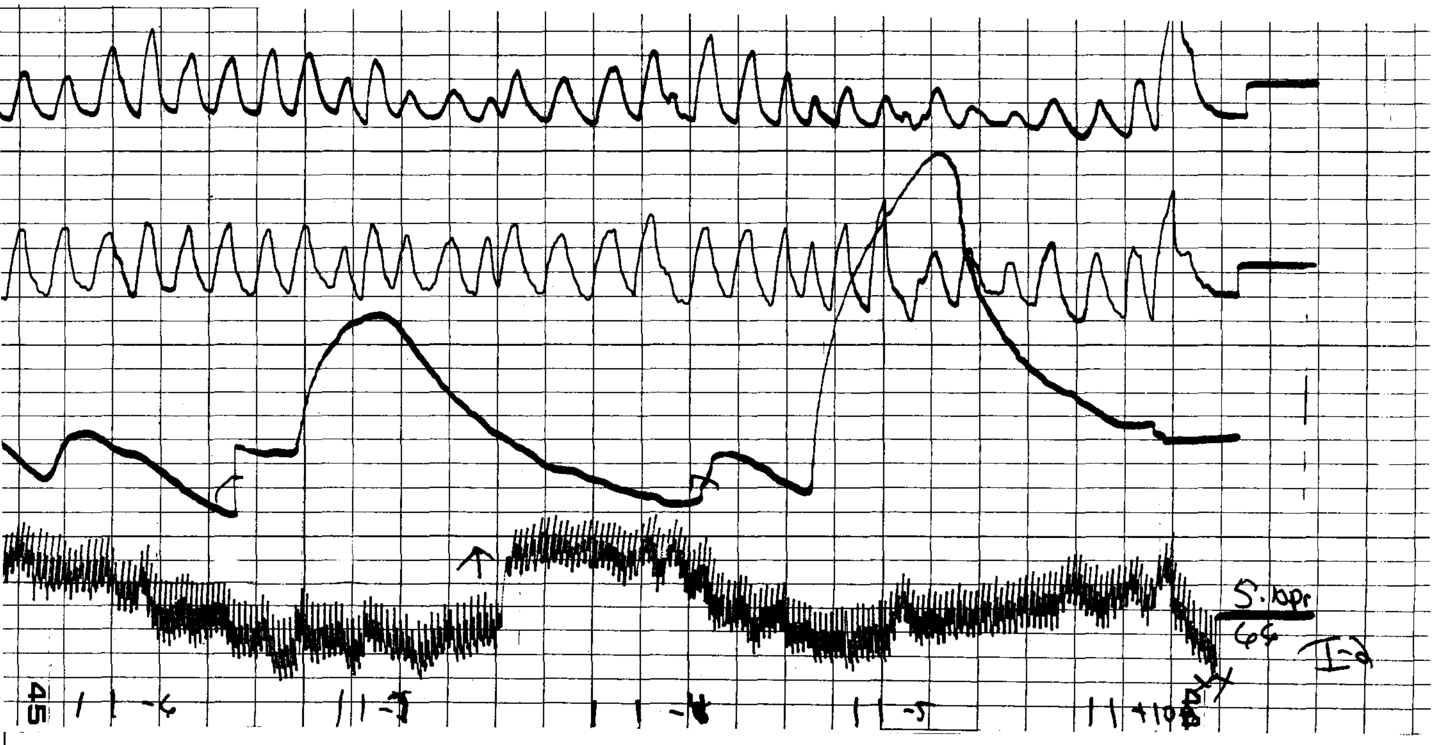


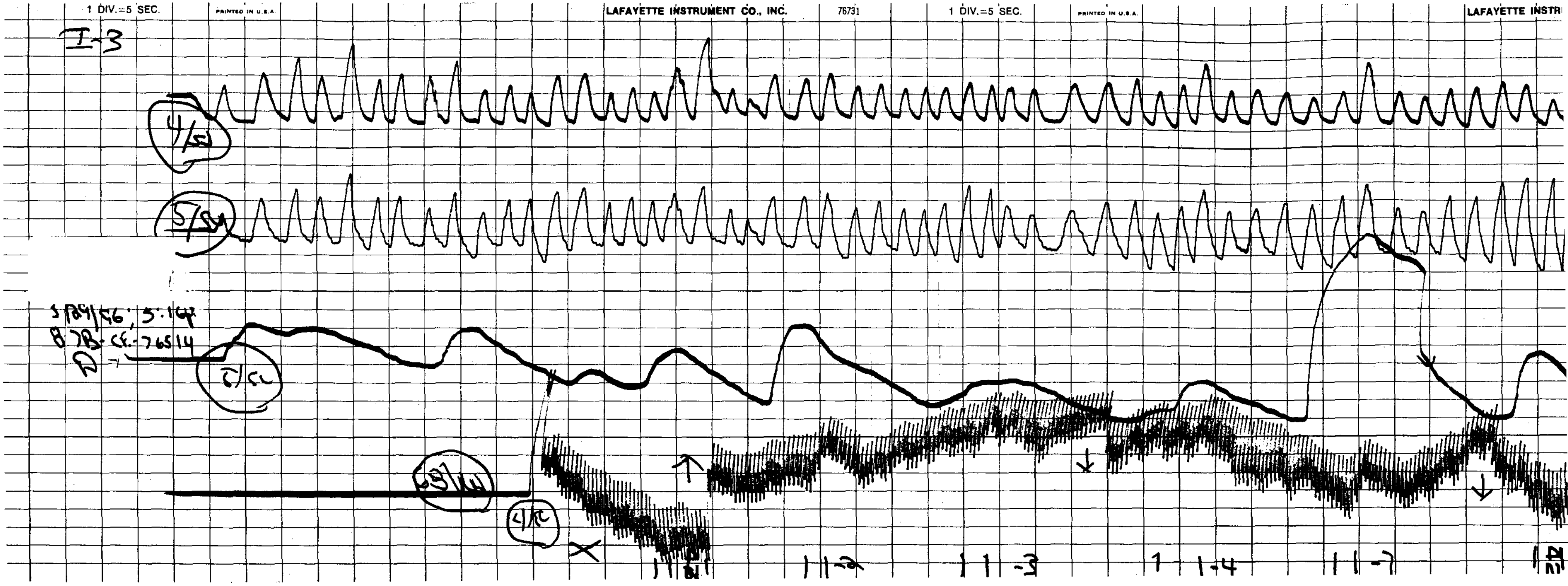
Identifi

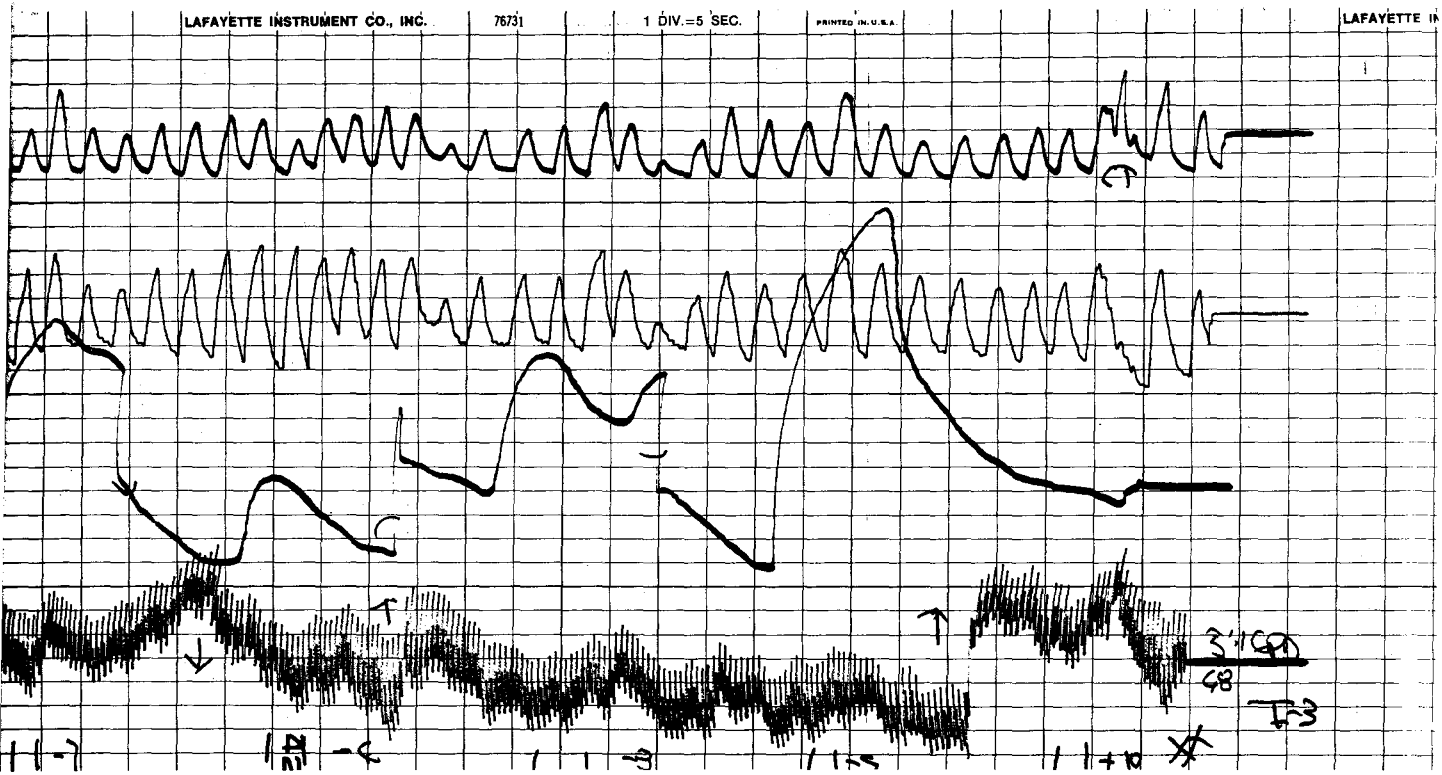

5-)
T
1 1-3
1 1-4
1 1-5
48
11-6
11-7
1 1-8
11-9


1 1-8
11-9
45
1 1-6
11-7
1 1-4
11-5


48
11-6
11-7
11-8
11-9
11-
47
5.01p
64
I-1


45
11-6
11-7
11-8
11-5
11+10
5.10pr
64
I-2


1 DIV.=5 SEC.
PRINTED IN U.S.A.
LAFAYETTE INSTRUMENT CO., INC.
76731
1 DIV.=5 SEC.
PRINTED IN U.S.A.
LAFAYETTE INSTR
I-3
4/52
5/54
3/29/56; 5:14P
8 JB-CE-76514
R
7/cc
8/74
4/RC
X
1 1 2
1 1-2
1 1 -3
1 1-4
1 1 -7
1 2


LAFAYETTE INSTRUMENT CO., INC.
76731
1 DIV.=5 SEC.
PRINTED IN U.S.A.
LAFAYETTE IN
CT
48
T-3
-7
-C
-3
+5
+10

Forensic Identification With Event Related Potentials

Vance MacLaren and Harald Taukulis

Abstract

The feasibility of using event-related brain potentials for detecting involvement in a simulated crime was explored. Sixteen participants either enacted a mock crime or merely read a description of the crime. They then were given a test in which true and false three-word sentences about the crime were flashed briefly on a computer monitor while the electroencephalogram was recorded. The first two words of each sentence were always presented for 500 milliseconds, followed by a white screen for 500ms. Imperative sentence completions were then shown for 500ms. Sixteen sentences were each presented 26 times in a random order. After each sentence, the participants said either "yes" or "no". Evoked P300 responses to the sentence completions and contingent negative variations in the interstimulus foreperiod were identified. Using a bootstrap index of P300 area to compare neural responses to statements answered with deceptive "no" responses versus nondeceptive but infrequent "yes" answers, conclusive test decisions were rendered in 12 cases. There was one false positive error, 5 true positives, and 6 true negatives. Detection using ERPs may develop into an effective and practical means of forensic identification.

Key words: brain wave, bootstrapping, event-related potentials, mock crime, N400, P300

Exploration of novel psychophysiological approaches to forensic identification is a continuing research effort (Yankee, 1995). Most of the existing research has focused on physiological response channels mediated by the sympathetic branch of the autonomic nervous system. Autonomic responses, such as skin conductance changes and respiratory depression, may be effective indicators of momentary changes in the level of stress being experienced by a suspect during questioning. However, in spite of nearly a century of active research, no autonomic response specific only to deception has been identified. Consequently, the effectiveness of truth verification interviews like the Comparison Question Technique (CQT; Reid, 1947) are largely dependent on the skill and experience of the person administering the test. For this reason, some academics have criticized practitioners who apply the CQT on the grounds that the test is not adequately standardized, prone to examiner bias, and difficult to administer in a way that is both fair and effective (e.g. Ben-Shakhar & Furedy, 1990; Lykken, 1998). While some forensic examiners may have great skill in applying the CQT, there can be no guarantee that all practitioners are equally competent. The requirement that the examiner be able to "sell" the interviewee on the subjective importance of the comparison questions is a particularly weak point, since there is no objective way of knowing the degree to which this has been accomplished in a given case. All of these problems could be circumvented if the observed physiological changes that occur in response to questioning could be related to a specific cognitive process. If that were possible, then an automated test could be developed that would obviate the argument that polygraph examinations are more art than science.

Author Note:

Special thanks to Ken Lerette for electronics support and to Kate Belyea for assistance throughout the project. Special thanks also to Donald Krapohl and Dr. Andrew Dollins for comments on an earlier version of the manuscript. This paper was presented at the August, 2000 meeting of the Canadian Association of Police Polygraphists in Ottawa. Address correspondence to Vance MacLaren, Department of Psychology, Hazen Hall, P.O. Box 5050, University of New Brunswick, Saint John, NB, Canada, E2L-4L5, vancem@nbnet.nb.ca.

One potentially fruitful avenue for the current research effort is the possibility that there may be phenomena that spontaneously occur within the central nervous system that are observable and could be used to discriminate a suspect's veracity. In recent years, a number of reports have surfaced detailing explorations of the feasibility of event-related potentials (ERPs) as the dependent measures in forensic detection tests [see Bashore & Rapp (1993) and Rosenfeld (1995) for reviews]. ERPs are indicators of processing in the central nervous system that are recorded using biopotential electrodes placed on the scalp that register changes in the electroencephalogram. Momentary changes in voltage are digitally sampled and averaged to give a representation of the activity of different brain areas when processing various types of stimuli that are presented to the subject. These measures have several advantages, including a high degree of specificity as indicators of identifiable cognitive processes, and relative automaticity and resistance to conscious manipulation by the subject.

In a majority of the studies of ERPs as detection indices (Allen & Iacono, 1997; Allen, Iacono & Danielson, 1992; Boaz, Perry, Raney, Fischler & Shuman, 1991; Farwell & Donchin, 1991; Pollina & Squires, 1998; Rosenfeld, Nasman, Whalen, Cantwell & Mazzeri, 1987; Rosenfeld, Cantwell, Nasman, Wojdac, Ivanov & Mazzeri, 1988), the aim has been to formulate and evaluate procedures that can detect the presence of concealed information. Such designs are analogous to the autonomic Guilty Knowledge Test (GKT; Lykken, 1959). In two additional studies (Johnson & Rosenfeld, 1992; Rosenfeld, Angell, Johnson & Qian, 1991), an attempt was made to create an ERP-based test that could screen for a variety of past infractions. Although these results have been impressive, guilty knowledge and pre-employment screening tests do not constitute the bulk of the work done by field examiners. It is far more common for present-day polygraph tests to be of the "specific issue" type and the GKT is notoriously difficult to apply under such conditions. In the present exploratory study, we examined the potential of an ERP-based technique designed to be applicable in situations that might otherwise see application of the traditional CQT. We applied this new test in two groups of participants: one group enacted a mock crime and the other did not, although both groups were fully informed about the simulated crime.

The present design replicates and extends previous work by Parsons (1996). Parsons' design is interesting because it was an attempt to capitalize upon both the P300 and N400 ERP components simultaneously in order to improve detection efficiency. The P300 (Sutton, Braren, Zubin & John, 1965) wave is reliably elicited by stimuli that are rare or that are in some way relevant to a task being performed by the subject (Johnson, 1986). The N400 wave is evoked by words that are anomalous in a given context (Kutas & Hillyard, 1980). With the growth in neuroscience that has occurred in the last few decades has come a better understanding of the conditions necessary for such ERP components to be observed. Both the P300 and N400 phenomena have been successfully applied as dependent measures in concealed information detection (e.g. Farwell & Donchin, 1991; Boaz, Perry, Raney, Fischler & Shuman, 1991). By manipulating the test conditions in such a way that both of these components occur differentially in guilty and innocent suspects, one might reasonably expect the resulting detection efficiency to be greater than if either component were manipulated in isolation.

In the present study, we followed the lead of Parsons (1996), Boaz, Perry, Raney, Fischler & Shuman (1991), and Stelmack, Houlihan & Doucet (1994) by presenting brief sentences to our participants that were comprised of separate sentence contexts (e.g. "Steven was...") followed by completions that were either correct (e.g. "shot") or not correct (e.g. "Stabbed"). It was predicted that correct completions would elicit P300 responses from participants in both the guilty and innocent groups when the sentence context referred to depersonalized facts about the crime. Such sentences are analogous to the comparison questions in a traditional CQT. It was also hypothesized that P300 responses to correct completions would also be observed in guilty participants when the contexts were self-referential (Ingram, 1995), but that this pattern would not be observed in the innocent group. In that sense, these sentences serve the same purpose as the crime-relevant questions in the CQT. The use of separate sentence contexts and completions was intended to capitalize on the N400 phenomenon, while the use of rare (true) and common (false) stimulus categories was intended to evoke processes concomitant with P300.

Methods

Participants

Seventeen English-speaking student volunteers were recruited from introductory psychology classes at the University of New Brunswick. For participating, they were offered two bonus points added to their grade. They were also told that they could earn a bonus of $5.00 contingent on their performance on the test, but all volunteers were given the cash prize. Data from one female guilty participant were eliminated from the sample because that individual had a skin condition that made electrode application difficult to the point that impedance at several of the recording sites could not be reduced to below 5 Kohms.

The mock crime guilty group consisted of 3 male and 5 female participants. Their ages ranged from 18 - 30 years (median = 19). Their WAIS IQ, as estimated using the Shipley Institute of Living Scale (Shipley, 1940; Zachary, Crumpton & Spiegel, 1985) ranged from 103 - 118 (median = 104). All had normal or corrected vision and reported no neurological impairment.

Four male and 4 female participants were assigned to the innocent group. Their ages ranged from 18 - 32 years (median = 20). Their estimated WAIS IQ ranged from 92 - 115 (median = 106). All had normal or corrected vision and reported no neurological impairment.

Apparatus

Electroencephalogram (EEG) was recorded continuously from the Fz, Cz, P3, Pz, and P4 scalp sites according to the International 10-20 system (Jasper, 1958). Recordings were made using a Grass model 8-10 electroencephalograph (www.grass-telefactor.com) and ECI electrode caps (www.electro-cap.com). Each channel was sampled at a rate of 500Hz by a Dataq Instruments (www.dataq.com) DI-220 12 bit analog to digital (A/D) convertor and stored on a personal computer for off-line analysis. All sites were referenced to linked earlobes. ECI tin biopotential electrodes were used at all sites and all electrode impedances were reduced to below 5 Kohm using ECI electrogel and mild abrasion. Impedance was checked immediately before and after each recording session. All EEG channels were band pass filtered between 0.1Hz and 35Hz, with an additional 60Hz notch filter to attenuate mains voltage contamination.

Artifacts in the EEG are often caused by eyeblinks. To monitor the occurrence of blinks and other eye movements, the electro-oculogram (EOG) was recorded using electrodes placed supraorbitally and below the outer canthus of the left eye. Because the EOG channel was to be used primarily for detection of eyeblink contamination, this channel was bandpass filtered between 1 Hz and 35Hz in order to reduce the oculomotor DC drift component. EEG channels were amplified using the 5 microvolt / millimeter sensitivity setting, and EOG was amplified to 10 microvolts / millimeter.

Recordings were made with the participants seated in an electrically shielded audiometric chamber with a glare resistant, flat computer monitor placed approximately 60cm from the participant at eye level. Throughout the recording sessions, participants were observed using closed circuit television and a microphone placed inside the chamber. Stimuli were presented on the computer monitor using Psychology Software Tools (www.pstnet.com) E-Prime (Beta 4.0) software. Sentence context and completion onset time points were marked with pulses sent via a serial port to a digital input channel of the A/D convertor. Digitized EEG, EOG, and event marks were stored on a second computer using Dataq Instruments Windaq software.

Procedure

All procedures were approved by the University of New Brunswick's ethical review committee. Volunteers read and completed a detailed consent form, and were then randomly assigned to either the mock crime enactment (guilty) or non-enactment (innocent) group. Both groups read details about the crime, but those in the guilty group were escorted by the experimenter to a store room to 'murder' a mannequin with a plastic model of a handgun which made a click when 'fired'. The guilty participants' instructions were as follows:

> A man named Steven owes you a lot of money and now it's payback time!
>
> Steven works for the university, but he is a very lazy employee and he likes to goof off. As usual, instead of doing his work, he has been drinking vodka and right now he is passed out in a storeroom. Sneak into the storeroom and look for a gun that he keeps hidden in there. Be careful not to wake him up! When you find the pistol, pick it up and murder Steven by firing a shot into his head. Put the revolver next to Steven's hand so it looks like he shot himself. After the murder, return to the testing room. When you return, you must not admit that you killed Steven under any circumstances. A murder like this could earn you a life sentence in prison!

Innocent participants remained in the laboratory after reading the following information about the crime:

> A man named Steven owed somebody a lot of money and he paid the ultimate price! Steven worked for the university, but he was a very lazy employee and he liked to goof off. As usual, instead of doing his work, he had been drinking vodka and was passed out in a storeroom. The assailant snuck into the storeroom and found a gun that Steven used to keep hidden in there. He was careful not to wake him up. The killer found the pistol, picked it up and murdered Steven by firing a shot into his head. The murderer then placed the revolver next to Steven's hand in an attempt to make it look as though he had shot himself.

After either enacting the simulated crime (guilty) or reading the description of it (innocent), all participants completed the Shipley Institute of Living Scale (Shipley, 1940). The Shipley scale is a quick index of verbal and abstract intellectual functioning known to give Intelligence Quotient (IQ) estimates comparable to those obtained by more elaborate intelligence tests like the WAIS-R (Zachary, Crumpton & Spiegel, 1985). This was done to screen for the possibility that intellectually impaired individuals might volunteer and subsequently respond to the test in an abnormal way because of their inability to understand the task. All of the participants had age-adjusted estimates of IQ that were higher than 85, which is one standard deviation below the population mean of 100. After completing the Shipley test, the recording devices were attached. All participants were verbally told by the experimenter that, "...I know that you are [guilty / innocent] and you know that you are [guilty / innocent], but the computer doesn't know. The purpose of the experiment is to see if this computer system can identify who is guilty and who is innocent of acting out the phony crime by analyzing their brain waves."

After being placed in the recording chamber, additional instructions were presented on the computer monitor. Specifically, they were told to read all of the sentences that were to appear on the screen, to decide whether each one was a true or false statement about the simulated crime, and to say "No." after each untrue statement. They were told to say "Yes." after true statements, but not to incriminate themselves by agreeing with any statements implying their guilt. They were also instructed to try to count the number of true statements and that if they could recall the correct number at the end of the test, they would receive a bonus of five dollars. This mental counting manipulation was intended to help ensure participant compliance with the instruction to attend carefully and to read all of the sentences. The recording session took approximately 21 minutes.

In some ERP studies, participants are required to respond with button presses, but we decided to use verbal responses because it was felt that this method would be more comfortable and understandable for the participants. Although verbal responding might introduce artifacts associated with muscular contractions in the throat and mouth, the latency of the verbal responses was expected to be longer than the ERP components of interest and were therefore not expected to contaminate the ERP results.

After completing the test and having the recording devices checked and removed, a five item recall test of details mentioned in the written descriptions of the crime was given to ensure that all participants remembered the crime at the time of their test. Participants were then paid, debriefed and dismissed. After completion of the experiment, a report of the results was mailed to each volunteer.

Materials

The test consisted of three word sentences presented on a computer monitor, with the last word appearing separately. For example, the sentence context "Steven was..." was followed by completions that were either correct ("shot") or not correct ("stabbed"). The sentence contexts were each presented for 500 milliseconds, followed by a blank monitor for 500 ms, and then the completion was shown for 500 ms. All stimuli were presented in a black 30 point Arial font on a white background and appeared at the center of the screen. Between trials the monitor remained white, without any words or a fixation point, for 1500 ms. ERPs were averaged for the period from 100 ms before the appearance of the sentence contexts through 1000 ms after onset of the completions. The following sentences were each presented 26 times in pseudorandom order:

I shot... / Steven
Steven was... / shot
I shot... / Kevin
Steven was... / stabbed
I strangled... / Steven
Kevin was... / shot
I strangled... / Paul
Kevin was... / clubbed
I stabbed... / Steven
Frank was... / shot
I stabbed... / Frank
Frank was... / strangled
I clubbed... / Steven
Peter was... / shot
I clubbed... / Peter
Peter was... / hanged

Analysis

The continuous EEG recorded from each channel was converted from arbitrary units into microvolts, parsed into epochs, sorted and averaged using a special macro written for SPSS (8.0) statistical analysis software (www.spss.com). Because eyeblinks create a strong electrical signal that can distort the recordings made at scalp sites, eyeblink contaminated data were eliminated from averaging on a point-by-point basis. Individual data points were omitted if they fell within a time window of 50 milliseconds before through 50 milliseconds after any other point having a value in the EOG channel that was either greater than 100 microvolts or less than -100 microvolts. Using this procedure, 14.74% of data points were lost due to EOG contamination, but no complete trials were omitted.

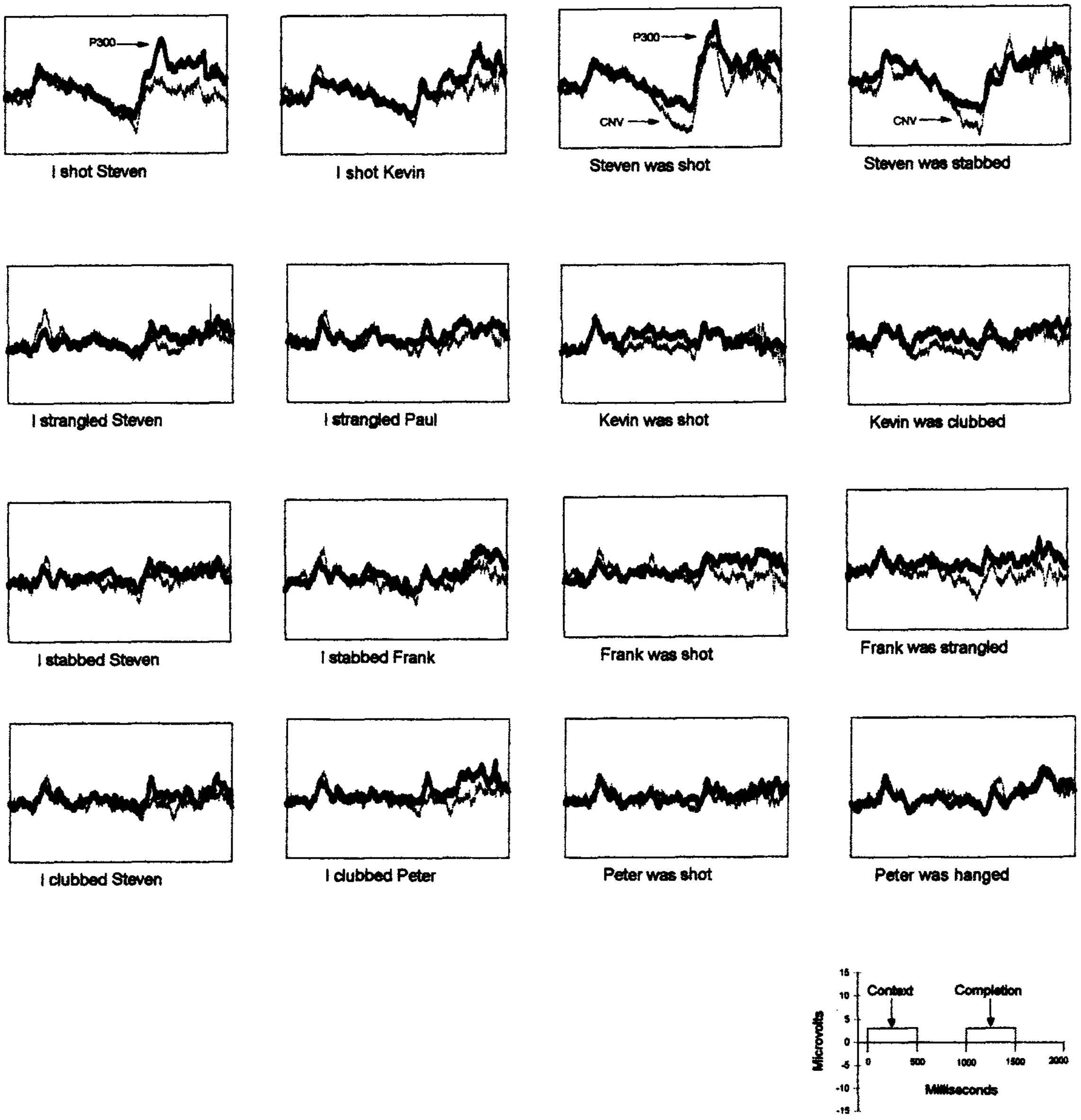
The P300 responses evoked by the sentence completions were the major dependent variable in the experiment. The amplitude of the P300 response was defined as the maximum voltage value occurring between 250ms and 600ms after the first appearance of the completion, minus a baseline. The baseline was calculated as the average of the 50 samples occurring in the 100ms immediately prior to the first appearance of the completion. P300 amplitude measurements were calculated using the averaged waveform for each participant at each of the five scalp sites (Fz, Cz, P3, Pz, and P4). Separate P300 amplitude measurements were computed for each of the 16 sentences.

Results

Grand Averages

Visual inspection of the grand averages depicted in Figure 1 reveals several distinguishing features. The ERPs are characterized by a small positive deflection occurring shortly after presentation of relevant sentence contexts, followed by a relatively large and slow negativity during the interstimulus foreperiod, and finally a large positive deflection following the correct sentence completions. These components are not apparent in the ERPs evoked by the non-relevant sentences. The group differences in waveforms evoked by the crime-relevant (ie "I shot... Steven.") and comparison sentences (ie "Steven was... shot.") suggest that the guilty subjects processed these two sentences in a very similar way, but those in the innocent group may not have. In particular, innocent participants' ERPs to comparison sentences showed greater negative deflection during the interstimulus interval and large post-completion positivity. Such differences are not apparent within the guilty group.

Figure 1. Grand average responses recorded at Cz from eight innocent (light line) and eight mock crime guilty (heavy line) participants.



I shot Steven

I shot Kevin

Steven was shot

Steven was stabbed

I strangled Steven

I strangled Paul

Kevin was shot

Kevin was clubbed

I stabbed Steven

I stabbed Frank

Frank was shot

Frank was strangled

I clubbed Steven

I clubbed Peter

Peter was shot

Peter was hanged

Inter-site Reliability

P300 amplitude measurements taken at the five scalp sites showed a high degree of redundancy, with significant (p<.01) correlations observed between all sites. The Pearson correlation coefficients ranged from .715 between P300 amplitude measurements taken at the P3 and Fz sites, to .960 between those taken at P3 and Pz. Because of this redundancy, subsequent analyses were focused primarily on the P300 responses measured at a single site, Cz. Measurements taken at this location were felt to be representative of those taken at the other sites. The Cz site is also the most practically relevant site, since it is the easiest one to physically locate and is therefore most likely to be used in future ERP-based field tests.

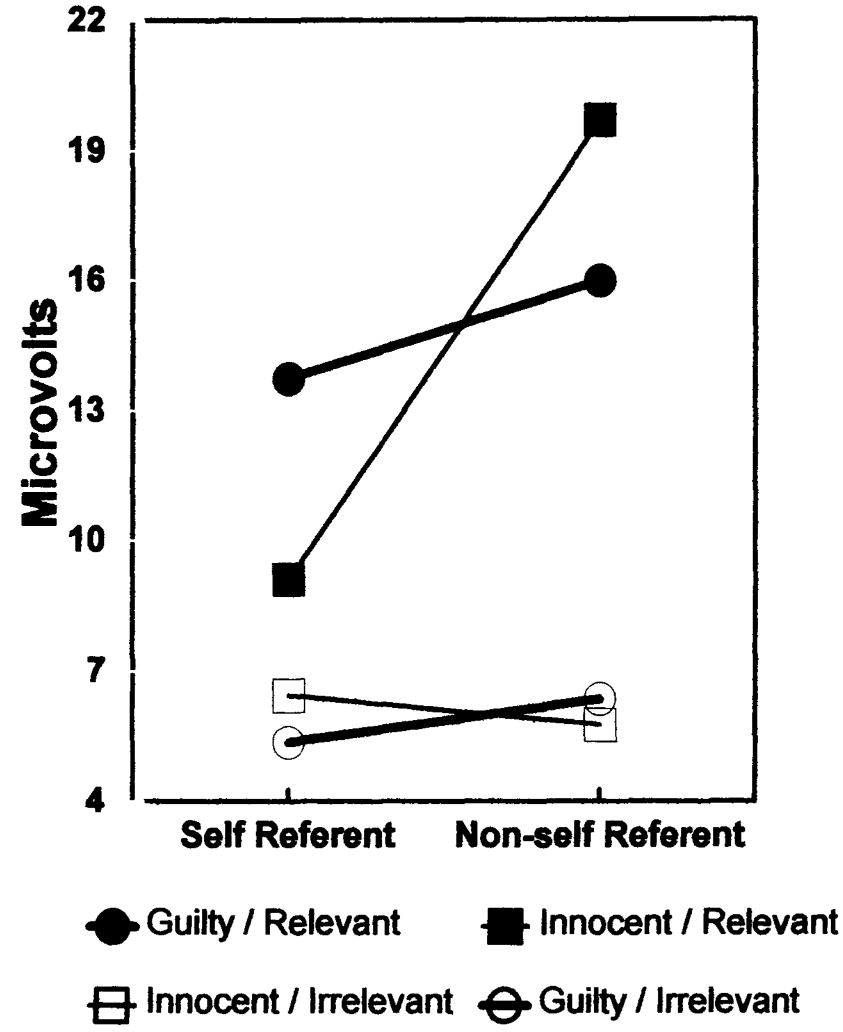
Group Effects

A multivariate analysis of variance (MANOVA) was conducted on the averaged P300 responses evoked by the sixteen sentences. The dependent variables were the P300 amplitude measures taken at the five scalp sites. For each subject, four independent fixed factors were coded for each of the sixteen sentences. These were the participant's guilt or innocence of enacting the mock crime, whether the sentence context was significant (e.g. "Steven was...") or irrelevant (e.g. "Frank was..."), whether the sentence was phrased in a self referent (e.g. "I stabbed...") or non-self referent way (e.g. "Kevin was..."), and whether the last word of the sentence was crime-related (e.g. "...shot") or a non-relevant control (e.g. "...strangled"). Significant third order interactions were found at each scalp site, with F(1,240) values ranging from 4.491 (p=.035) at Pz to 9.292 (p=.003) at Fz,. The interaction was significant (alpha = .05) at all five sites, including Cz (F(1,240) = 4.73, p = .031). This interaction involved guilt versus innocence, the relevance of the sentence context, and the self-referential phrasing of the sentence. The interaction is depicted graphically in figure 2.

In post hoc tests it was found that, across the guilty and innocent groups, participants had significantly (t(254) = 10.50, p<.001) larger Cz P300 responses evoked by sentences with relevant (mean = 14.62, SD = 8.98) contexts than by sentences with irrelevant contexts (mean = 5.99, SD = 4.06). Similar differences at the other scalp sites were also significant at the alpha = .001 level. Innocent participants' P300 responses to sentences with relevant contexts were particularly large (t(30) = 3.09, p=.004) when the sentence was non-self referent (mean = 19.68, SD = 8.99), as compared to when it was self referent (mean = 9.10, SD = 10.37). These differences were also significant at the other scalp sites, with alpha = .05. Among guilty participants, Cz P300 responses to sentences with relevant contexts did not differ significantly (t(30) = .96, p=.344) between sentences that were non-self referent (mean = 15.98, SD= 7.13) and those that were self referent (mean = 13.73, SD = 16.03). Such differences at the other scalp sites also failed to reach significance, with alpha = .05.

The crime-relevant ("I shot... Steven") and comparison ("Steven was... shot") statements used in this study are equivalent in terms of their information content. To a deceptive guilty subject, the two statements are essentially identical, but they might be perceived as being very different by a nondeceptive innocent subject. Because they said "Yes" in response to "Steven was... shot" and "No" to all other sentences, we expected to find large P300 responses in the comparison sentence. Because only guilty subjects were deceptive when saying "No" after "I shot... Steven", then the presence of a P300 response would indicate deception-related significance. Among guilty subjects, P300 responses were expected to accompany both sentences, but to be triggered only by comparison sentences in the innocents. Therefore, within-subject differences in physiological responding to those two sentences could be reasonably attributed to processes associated with deception. It was therefore decided to carry out planned comparisons of the guilty and innocent subjects' P300 responses to these two sentences. Innocent participants' Cz P300 responses to the comparison sentence were significantly greater (mean difference = 8.6 microvolts (SD = 7.23), paired t(7) = 3.37, p = .012) than their responses to the crime-relevant sentence. Among guilty participants, the Cz P300 responses to the comparison sentence were not significantly different (mean difference = 1.25 microvolts (SD = 6.32), paired t(7) = 0.56, p = .594) than their responses to the mock crime-relevant sentence.

Figure 2. Interaction of guilt versus innocence, self referent versus non-self referent sentence contexts, and crime relevant versus non-relevant completions on amplitude of P300 responses recorded at Cz.



Exploratory Analysis of Group Differences

Because this was an exploratory study, it seemed prudent to consider the possibility that unanticipated group differences might be present. Average difference waveforms were constructed by subtracting each data point in the ERP waveform evoked by the crime-relevant sentence from the corresponding points in the comparison sentence waveform for each subject. This yielded a plot of the average difference between each subject's responses to the two sentences. These difference waveforms were then parsed into 20 consecutive sub-epochs, each lasting 100ms. The area under each sub-epoch was calculated by adding a constant of 100 to each data point, summing the 50 points, then subtracting 5000 from the total. This process was repeated for each subject and each scalp site. The areas were then used as the dependent measures in a multivariate analysis of variance, with guilt versus innocence as a fixed factor. In this analysis, the only variable that was found to differ significantly (alpha = .05) was the difference waveform area in the interval from 1000ms - 1100ms after onset of the sentence context. These differences were significant at the Cz ($\underline{F}(1,14) = 9.63$, $\underline{p}=.008$) site. The negative ERP component can be clearly seen in the grand averages and is indicated by the term "CNV" in Figure 1.

Individual Diagnoses

To make intra-subject diagnoses of guilt or innocence, we followed the lead of Farwell & Donchin (1991) and used a non-parametric statistical technique known as bootstrapping (Wasserman & Brockenholt, 1989; see also Honts & Devitt, 1992). Because amplitude measures would likely be highly contaminated by error variance, P300 area at Cz was used as the discriminative variable. An estimate of P300 area was calculated for each trial, and the area values were used in the bootstrapping procedure.

To reduce the influence of trial-to-trial variations in baseline, the average of the 50 samples (100ms) immediately preceding the onset of each sentence completion was subtracted from every data point recorded thereafter. In cases where data in the baseline period were missing due to EOG artifact, a value of zero was used as a substitute. The area under the EEG waveform in the period from 250ms through 600ms after the presentation of each sentence completion was calculated. To do this, a constant of 100 was added to each of the 175 data points, the values were summed, and 17,500 was subtracted from the total. For each subject, this area calculation was done for all 26 of the relevant sentence presentations and the 26 comparison sentences. To compute the P300 area difference, the areas corresponding to each of the 26 relevant sentences were averaged and the areas corresponding to each of the 26 comparison sentences were averaged. These two averages were subtracted to yield the observed P300 area difference between the relevant and comparison statements.

The probability of obtaining the observed P300 area difference by chance, if the two areas were actually equivalent, was estimated by iterative bootstrapping using Resampling Stats (4.2) software (www.resample.com). Areas under the 26 relevant and 26 comparison trials were 'concatenated' into a single data file and 'shuffled' using Resampling Stats 4.02. A distribution of expected area differences for each subject was created by taking 10,000 random samples from the combined data file. In each estimate, 26 areas were randomly selected, with replacement, to represent a hypothetical set of relevant epochs and 26 were selected to represent the comparison epochs. Each of these epoch sets was averaged and the difference between the two averages was calculated. By repeating this process 10,000 times, a 'bootstrap distribution' of the expected area difference was created for each subject. The actual P300 area difference that was observed in each individual was compared against that individual's bootstrap distribution. The likelihood of obtaining an area difference equal to, or greater than, the observed area difference was estimated by counting the number of hypothetical area differences in the bootstrap distribution that were equal to, or greater than, the observed difference. By dividing this number by 10,000, the probability of obtaining the observed area difference was estimated, and given as a bootstrap probability index (p).

The p value is an estimate representing the likelihood that the P300 area difference observed in the individual could occur by chance if his or her actual P300 responses to the relevant and comparison questions were really equivalent to one another in terms of area. The p values are inversely related to the magnitude of the difference between P300 responses to relevant and comparison statements; lower p values indicate a greater chance that the subject processed the two statements differently.

To arrive at individual decisions about guilt, cutoff scores were used to classify individual participants as truthful, inconclusive or deceptive, based on their bootstrap probability index. We used the same cutoff points used by Farwell & Donchin (1991) and Allen & Iacono (1997), namely .10 and .70. Participants with p scores less than .10 were classed as truthful, those with scores above .70 were classed as deceptive, and any others were called inconclusive. Using these criteria, 6 innocent and 5 guilty subjects were correctly classified. There were 4 inconclusive outcomes and 1 false positive (see Table 1.). Of the 12 cases in which conclusive outcomes were obtained, 11 (91.6%) were correct.

It is worth noting that the false positive error was in the case of an innocent subject who apparently misunderstood the procedural instructions. When reviewing the videotape recording of her session, it was noticed that she said "Yes." in response to both "Steven was shot" and to "I shot Steven". It is therefore not surprising that anomalous results were obtained in that individual.

Discussion

The results of this study supported the major hypothesis that P300 amplitude would differ between crime-relevant and comparison statements among innocent, but not guilty participants. These group differences were sufficient to allow individual diagnoses to be made with a reasonable degree of accuracy. In addition to formulating a field applicable form of an ERP-based technique for detecting deception, further research should be aimed at understanding the mechanism underlying the pattern of P300 responses described here.

In addition to the P300 results, we also noted a slow negative component that occurred near the end of the interstimulus foreperiod. We have tentatively identified this negativity as a Contingent Negative Variation (CNV; Walter, Cooper, Aldridge, McCallum & Winter, 1964). The CNV may be treated as a physiological manifestation of processes related to anticipation or expectancy of an imminent stimulus that requires either motor reaction or cognitive processing (Damen & Brunia, 1994). The CNV and P300 are intimately related phenomena. As argued by Verleger (1988; see also Deeke & Lang, 1988), the P300 occurring at the end of a perceptual epoch may be the result of the "resetting" of expectancy-related negativity that occurs within the epoch. In the present experiment, participants in the guilty group showed equally large P300 responses to the crime-relevant (i.e. "I shot... Steven") and the comparison (i.e. "Steven was... shot") statements. The innocent participants, however, displayed both CNV and P300 responses that were not equivalent in size. Previous studies by Parsons (1996) and by Boaz, Perry, Raney, Fischler & Shuman (1991) also found large CNV shifts in the interval between sentence contexts and completions. Additionally, our lab previously reported that acquired expectancies and violation of such expectancies can have powerful effects on electrodermal detection in the GKT (MacLaren & Bradley, 1998). The presence of differences, both in CNV and P300, is consistent with our earlier findings with electrodermal responses.

Among members of the guilty group, the absence of within-subject differences in CNV and P300 may be explained using notions of priming and expectancy. According to both the triarchic model of P300 (Johnson, 1986) and the context updating hypothesis (Donchin & Coles, 1988), the perceived meaning of a stimulus can be a key determinant of the magnitude of the P300 response that it elicits. For participants in the guilty group, both "Steven was..." and "I shot..." may have served as effective primes for the expectation that the imperative sentence completions could be subjectively meaningful. In the former case, the word "shot" would indicate the need to say the word "Yes", which was a rare behavioral response. In the latter, to say "No" after "Steven" was a common answer, but it was also a rare lie amongst common truthful answers. In both cases, the sentence had a 50% chance of being completed by a word that would lend significance to the statement.

The expectancy model can also account for the within-subject differences observed among innocent subjects. Since only the "Steven was..." context could have been completed by a significant ending, people in the innocent group may have been primed to expect significant endings only after "Steven was..." and to disregard the "I shot..." sentences as non-significant. It is possible that the innocent subjects decided that any words completing the "I shot..." statements would be non-significant, perhaps even before they were presented. Because of this priming effect, their expectancy-related CNV and any significance-related P300 responses could have been blocked. According to this explanation, the neural mechanisms associated with significance-related P300 responses may be 'switched' on or off by the sentence contexts. When switched off, preparatory processes manifested by CNV are also not seen.

If reliable, the present results could lead to the development of an ERP-based technique that could be used in situations now handled using the CQT. By exploiting priming sentence contexts as a way of manipulating the way in which critical information is presented to the subject, physiological responses might be brought under experimental control. In this study, we found that guilty participants sustained their perception of the crime-relevant statements as being important enough to elicit P300 responses. The priming effect sufficiently reduced innocent subjects' P300 responses so that false positive errors were avoided in all but one case, even though all innocent subjects were exposed to the same crime-relevant information as members of the guilty group. Further testing will be required to see if the accuracy of this ERP-based method of forensic detection can be as accurate as the autonomic CQT.

Table 1. Bootstrap indices and test outcomes.

Subject ID	Bootstrap Index	Classification
Guilty #1	.756	Correct
Guilty #2	.995	Correct
Guilty #3	.860	Correct
Guilty #4	.295	Inconclusive
Guilty #5	.132	Inconclusive
Guilty #6	.976	Correct
Guilty #7	.944	Correct
Guilty #8	.235	Inconclusive
Innocent #1	.000	Correct
Innocent #2	.002	Correct
Innocent #3	.066	Correct
Innocent #4	.013	Correct
Innocent #5	.983	False Positive
Innocent #6	.328	Inconclusive
Innocent #7	.001	Correct
Innocent #8	.004	Correct

Although the present results are encouraging, this technique is in a primitive stage of development. Until the results can be independently replicated and an optimal form of the test developed, no application is warranted. The present sample was small and the volunteers in this study may not be representative of actual criminal suspects. Also, the mock crime paradigm used in this study is highly artificial and bears only a superficial resemblance to the conditions of a real investigation. Nevertheless, the magnitude of the observed effects were large enough to reach statistical significance, despite a low level of statistical power for hypothesis testing. We are therefore optimistic that these results might be replicable in other labs and perhaps generalize to non-simulation conditions. It is our hope that these findings may stimulate others to pursue a more technologically sophisticated approach to forensic psychophysiology.

References

Allen, J.J.B. & Iacono, W.G. (1997). A comparison of methods for the analysis of event-related potentials in deception detection. *Psychophysiology,* 34, 234-240.

Allen, J.J., Iacono, W.G. & Danielson, K.D. (1992). The identification of concealed memories using the event-related potential and implicit behavioral measures: A methodology for prediction in the face of individual differences. *Psychophysiology,* 29, 504-522.

Bashore, T.R. & Rapp, P.E. (1993). Are there alternatives to traditional polygraph procedures? *Psychological Bulletin,* 113, 3-22.

Ben-Shakhar, B. & Furedy, J.J. (1990). *Theories and Applications in the Detection of Deception.* New York: Springer-Verlag.

Boaz, T.L., Perry, N.W., Raney, G., Fischler, I.S. & Shuman, D. (1991). Detection of guilty knowledge with event-related potentials. *Journal of Applied Psychology,* 76, 788-795.

Damen, E.J.P. & Brunia, C.H.M. (1994). Is a stimulus conveying task-relevant information a sufficient condition to elicit a stimulus-preceding negativity? *Psychophysiology,* 31, 129-139.

Deeke, L. & Lang, W. (1988). P300 as the resolution of negative cortical DC shifts. *Behavioral and Brain Sciences,* 11, 379-381.

Donchin, E. & Coles, M.G.H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences,* 11, 357-374.

Farwell, L.A. & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology,* 28, 531-547.

Honts, C.R. & Devitt, M.K. (1992). Bootstrap Decision Making for Polygraph Examinations. Fort McClellan, AL: Department of Defense Polygraph Institute (N00014-92-J-1794).

Ingram, E.M. (1995). Event-Related Potentials: The P300 and Self-Referent Stimuli. Ft. McClellan, AL: Department of Defense Polygraph Institute (DoDPI94-R-0006).

Jasper, H.H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology,* 10, 371-375.

Johnson, R. (1986). A triarchic model of P300 amplitude. *Psychophysiology,* 23, 367-384.

Johnson, M.M. & Rosenfeld, J.P. (1992). Oddball-evoked P300-based method of deception detection in the laboratory II: Utilization of non-selective activation of relevant knowledge. *International Journal of Psychophysiology,* 12, 289-306.

Kutas, M. & Hillyard (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science,* 207, 203-205.

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

Lykken, D.T. (1998). *A Tremor in the Bood.* New York: Plenum.

MacLaren, V.V. & Bradley, M.T. (1998). Conditioning of expectations in a concealed knowledge test. *Polygraph,* 27, 157-170.

Parsons, T.E. (1996). Event-related potentials and the detection of guilty knowledge. Unpublished Doctoral Dissertation, University of Georgia.

Pollina, D.A. & Squires, N.K. (1998). Many-valued logic and event-related potentials. *Brain and Language,* 63, 321-345.

Reid, J.E. (1947). A revised questioning technique in lie-detection tests. *Journal of Criminal Law and Criminology,* 37, 542-547.

Rosenfeld, J.P. (1995). Alternative views of Bashore & Rapp's (1993) alternatives to traditional polygraphy: A critique. *Psychological Bulletin,* 117, 159-166.

Rosenfeld, J.P., Angell, A., Johnson, M. & Qian, J. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average waveforms. *Psychophysiology,* 28, 319-335.

Rosenfeld, J.P., Cantwell, B., Nasman, V.T., Wodjac, V., Ivanov, S. & Mazzeri, L. (1988). A modified, event-related potential based guilty knowledge test. *International Journal of Neuroscience,* 42, 157-161.

Rosenfeld, J.P., Nasman, V.T., Whalen, R., Cantwell, B. & Mazzeri, L. (1987). Late vertex positivity as a guilty knowledge indicator: A new method of lie detection. *International Journal of Neuroscience,* 34, 125-129.

Shipley, W.C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *Journal of Psychology,* 9, 371-377.

Stelmack, R.M., Houlihan, M. & Doucet, C. (1994). Event-Related Potentials and the Detection of Deception: A Two-Stimulus Paradigm. Fort McClellan, AL: Department of Defense Polygraph Institute (N0001493-1-1002).

Sutton, S. Braren, M., Zubin, J., & John, E.R. (1965). Evoked potential correlates of stimulus uncertainty. *Science,* 150, 1187-1188.

Verleger, R. (1988). Event-related potentials and cognition: A critique of the context updating hypothesis and an alternative interpretation of P3. *Behavioral and Brain Sciences,* 11, 343-356.

Walter, W.G., Cooper, R., Aldridge, V.J., McCallum, W.C. & Winter, A.L. (1964). Contingent negative variation: A electric sign of sensorimotor association and expectancy in the human brain. *Nature,* 203, 380-384.

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

Yankee, W.J. (1995). The current status of research in forensic psychophysiology and its application in the psychophysiological detection of deception. *Journal of Forensic Sciences*, 40, 63-68.

Zachary, R.A., Crumpton, E. & Spiegel, D.E. (1985). Estimating WAIS-R IQ from the Shipley Institute of Living Scale. *Journal of Clinical Psychology*, 41, 532-540.

Polygraph Validity in the New Millennium

Stan Abrams

Abstract

This is an investigation into various aspects of polygraph validity and reliability that can be divided into five segments. The major portion of this paper deals with the evaluation of 100 confirmed truthful and deceptive single-issue computerized polygraph charts that were blind scored independently by two polygraphists. Employing 7- and 3-position scales with various cutoff points, the findings indicated that the 7-position scale with a +/-6 cut off had the highest accuracy, but also one of the highest inconclusive rates. In contrast to prior research, accuracy was higher for the truthful than the deceptive. Comparisons were made with the validity findings of Patrick and Iacono (1987, 1991) and the original examiners and the blind evaluators. It was determined that the accuracy of the original examiner is always higher, but there is good reason to believe that it is because they use extrapolygraphic information not available to the blind evaluator. In the second part of this study, various indices of reactivity were evaluated, to determine which of these were the most effective in evaluating truthfulness and deception. Particular emphasis was placed on respiration and some of the Department of Defense Polygraph Institute's (DoDPI) recommended 22 evaluative criteria (Ansley & Krapohl, 2000). A number of the latter were not found to be of value in polygraph chart evaluations either because they occurred too rarely or because they occurred almost equally on either the control or relevant so that their accuracy was strictly a matter of chance. In the third phase of this investigation the various sensors were compared to determine which provided the most accurate data for a final determination of truth or deception. As has been reported in most research, the electrodermal response contributed the most followed by the cardiograph and finally respiration. In the fourth portion of this work, a study was made of the validity of Patrick and Iacono's argument that one cannot generalize from confirmed polygraph findings that were based on admissions to unconfirmed examinations such as those being considered for admissibility into court (Patrick & Iacono, 1991). They hypothesized that those examinations that were confirmed by confessions were quite different because they were likely to have higher deceptive scores, thereby motivating the examiner to interrogate to a greater degree. Because of that more admissions would be made. On the other hand, they argued that those with lesser scores would probably include truthful subjects who would not make admissions. To determine if these were accurate statements, scores of deceptive confirmed and unconfirmed charts were compared, and in addition, 24 unconfirmed cases in which deception was found were tracked through the court system. These results contradicted Patrick and Iacono. The final phase of this study attempted to evaluate whether scoring to the stronger comparison question on one relevant question, which served to reduce false positive errors, could be expanded to two relevant questions without increasing the risk of false negative errors. The findings which were determined on a 3-position scale using a +/-3 cutoff demonstrated about a 5% risk of false negative errors. This could have been eliminated by using a cutoff of +/-4.

Key words: 3-position scoring, 7-position scoring, field cases, reliability, validity

For reprints, contact Stan Abrams, Ph.D., P.O. Box 69571, Portland, Oregon 97201

Background of Polygraph Validity

In the over 100 years in which polygraphy has been in existence, there has been a great deal of research into its validity. However, in 1983 the Office of Technology Assessment (OTA, 1983) reported that the vast majority of the research that had been conducted was inadequate. For their study of the validity of polygraphy at the request of the United States Congress, only ten studies met their criteria for inclusion into their research. Seven of these studies dealt with the blind scoring of polygraph tests. Their research indicated a rather large range of accuracy for the various studies but showed a mean validity for the deceptive of 86.3 % and 76.0 % for the truthful. Unfortunately this investigation had its own weaknesses, one of which was associated with the risk of combining politics and science. Those who were conducting this investigation were well aware of Congress' interest, which was to satisfy the wishes of the unions and ACLU through the elimination of polygraphy from the workplace. This could have created a considerable bias. Other problems associated with this research included scoring inconclusives as errors and including one research paper by Kleinmuntz and Szucko (1982) that was significantly flawed in its design because student examiners were used, no inconclusive decisions were permitted, and only one polygraph chart was utilized.

As will be seen, one of the major problems associated with blind scoring is that high levels of inaccuracy can be spuriously obtained when the original examiners employed extrapolygraphic information such as case facts and an evaluation of subject behavior in making their decisions. Related to this, one of their findings was that specific polygraph testing yielded very poor results with truthful subjects. They pointed out that this was true in the research of Horvath, (1977), Raskin (1978), Barland & Raskin (1976) and Kleinmuntz and Szucko (1982), all of whom employed blind scoring techniques.

Patrick and Iacono's (1987) study evaluated the findings of police examiners and reported 100% accuracy with the deceptive and 90% with truthful subjects. In their later research they stated that one couldn't rely on the polygraph findings reported by the original examiners because their decisions were influenced by extrapolygraphic information. (Patrick & Iacono, 1991) They further indicated that only in the blind scoring of charts can one determine the true validity of polygraphy. When these charts were re-scored blindly the results were 98% accuracy for the deceptive but only 55% percent for the truthful. It should be noted that the polygraph tests that were used in this investigation were administered between 1980 and 1984; therefore, they did not score to the stronger comparison question. In addition their study included multiple-issue tests and the -3 or less DoDPI methodology which resulted in a finding of deception in any single relevant question received a -3 or less. Inevitably this resulted in increasing the false positive findings. Both the OTA (1983) library research and Patrick and Iacono's investigation have had a profound and lasting negative impact on polygraphy.

Since that time there has been considerable research in which high levels of validity have been reported from all parts of the world. Many of these studies can be found in *The Validity and Reliability of Polygraph Testing* in Volume 26 of this journal (1997). For the 12 studies conducted since 1980 of confirmed field polygraph examinations of the original examiners, 96% accuracy was found for truthful subjects and 98% for those who were deceptive. Blind scoring research on confirmed field tests was 90% accuracy for the truthful and 95% for the deceptive.

Swinford (1999) published the 22 DoDPI rules for scoring the 7-position scale and more recently Ansley and Krapohl (2000) evaluated the frequency of the indices of reaction. The top five, that is, those that appeared at least 5% of the time were: amplitude change (26%) and duration (24%), both in the electrodermal sensor; baseline change, that is, an increase and then a decrease in the pneumograph (15%); complexity in the electrodermal (6%), and; amplitude decrease in the cardiograph (51%). Premature ventricular contractions (PVCs) or irregular heartbeats were not included in the 22 indices of reaction because their frequency of occurrence was too low.

More research has been devoted to determining how much each sensor contributes to the total score. Ansley and Krapohl (2000) in their paper also reported that electrodermal produced 55% of the reactions, cardiograph 26%, and pneumograph 19%. Findings in the same direction were found by Waid et al. (1981), Bradley and Janisse (1981), Ohnishi (1976), and Abrams (1987).

Capps and Ansley (1992) reported that scoring confirmed truthful charts to the stronger comparison question for the first relevant question resulted in a greater rate of accuracy and fewer inconclusives than scoring to the weaker comparison question. With confirmed deceptive charts, the stronger comparison approach was only slightly less accurate than the weaker comparison method, but there were a larger number of inconclusives. Abrams (1997) replicated this study employing tests that had been scored completely to the preceding comparison question and then they were re-scored so that a comparison could be made among the stronger, weaker, and preceding techniques. The stronger comparison method only applied to the first relevant question, while the preceding comparison technique, of course, was applied to both relevants. In the case of the weaker comparison questions, Backster's method was used with both relevants being compared to the weaker comparison questions. For the confirmed truthful charts, there was complete accuracy for the stronger and preceding comparison approaches, but in the case of the weaker comparison question there were only 90% correct decisions. The inconclusive rate for both the stronger and preceding methods was 0, but it was 40% for the weaker comparison question approach. In the confirmed deceptive cases, there was complete accuracy across the board, but there was a 10% inconclusive rate for the stronger comparison technique. The findings were in close agreement with those of Ansley and Capps (1992). The stronger comparison question approach successfully reduced the false positive error rate without significantly affecting the false negative rate. It was felt that the preceding comparison question method was spuriously high because all of the charts had previously been scored in that manner. While the weaker comparison technique was particularly effective with deceptive subjects, there was a 10% error rate and 40% inconclusive rate with the truthful subjects.

The average scores for the deceptive charts were -12 for the stronger comparison, -15 for the preceding, and -18 for the weaker comparison. In contrast to those findings on the truthful charts it was +15, +11, and +6 respectively.

Method

One hundred polygraph charts, 50 confirmed truthful and 50 confirmed deceptive were used in this study. They all had been field conducted by federal or law enforcement examiners employing Axciton computerized polygraphs. All of the examinations were single-issue tests using three relevant questions. Most of these examinations used a weak relevant question at the third spot, such as, "Do you know who committed this act?" For these writers, these are seen as basically multiple-issue tests that can reduce the accuracy of the results. It was later learned that one of the examinations was a multiple-issue test and was eliminated from the sample, leaving 99 cases. The investigators in this study had no knowledge of the case facts, the questions asked, or the subject's behavior. It was assumed that the charts were scored by the original polygraph examiners using the 7-position scale with a +/-6 cutoff since these are the conventions in federal and law enforcement settings from where this sample was collected.

Each of the two evaluators independently scored the 99 charts using both the 7-position and 3-position scales. The cutoff points for the 7-position scale were +/-6, +/-4, -6/+4, the DoDPI method and its variations. "Full DoDPI" specifically refers to a -3 or less at any spot or a total of at least -6 for all spots to be considered a deceptive chart. DoDPI requirements for a truthful chart are a positive score in every spot and a total of +6 or greater. These relate to the DoDPI tri-spot zone comparison technique (Matte 1996).

For the 3-position scale the comparisons were made for cutoffs of +/-2, +/-3 and +/-4. False positive and false negative decisions were determined for each of these areas as well as the inconclusive rate and the degree of agreement between the two examiners (reliability). These results can be seen in Table 1. Various indices of reaction were totaled for the relevants on the confirmed deceptive charts and for the comparison questions on the truthful tests. A determination was then made to ascertain what percentage of time they were of value in determining truth or deception or whether their appearance was simply a chance occurrence that could just as readily be in the accurate direction as the inaccurate. Those criteria that were found to be of value were recommended for use, in contrast to those others that were clearly of little worth.

Many studies have been conducted to determine which sensors contributed the most in making accurate polygraph decisions. This was evaluated by totaling those scores for each sensor that were in the appropriate direction, for example, on the comparison question on confirmed truthful charts and on the relevant question on confirmed deceptive tests. Employing a second approach, in those cases in which the reactions were on the inappropriate questions, these results were subtracted from the accurate scores. In this manner, a determination was made for the contribution of each sensor.

Table 1. Validity of 3- and 7-position scales compared with various cutoff points
(Percentages are rounded off to the nearest whole number)

	7-position scale					3-position scale		
	+/-6	+/-6 or-3	Full DoDPI	+/-4	-6/+4	+/-3	+/-4	+/-2
Reliability Excluding Inconclusives	1.00	0.91	0.96	0.99	1.00	0.93	0.98	0.94
Average False Positive	1.5	7.5	7.5	2.5	1.5	4	3	4.5
Average False Negative	4	3.5	1	6.5	65	7	5.5	7
Correct Deceptives	89%	93%	98%	85%	83%	84%	86%	85%
Correct Truthful	97%	84%	76%	95%	97%	92%	94%	92%
Total Accuracy	93%	87%	88%	90%	90%	87%	89%	87%

To test Patrick and Iacono's (1991) hypothesis that a difference existed between confirmed and unconfirmed deceptive charts, a random sample of 20 of each were drawn and their total scores calculated. In addition, 24 unconfirmed deceptive charts were tracked through the court system to determine the degree of agreement with the deceptive test findings.

Finally, an attempt was made to determine if employing scoring to the stronger comparison questions on two relevant questions, which would further reduce false positive errors, would result in an increase in false negative findings. Twenty tests that had been previously found truthful and twenty that had been evaluated as inconclusive were randomly selected. Each had been previously scored to the stronger comparison question on both relevants using the 3-position scale with a cutoff of +/-3. They were re-evaluated by scoring the second relevant to the preceding comparison question, thereby, moving the total scores in the direction of deception. A determination was made as to how many of these cases then fell into the deceptive range.

Results

It must be recognized that when the cutoff points to determine truth or deception are increased, the inconclusive rate increases as well, and of course, the opposite occurs if the cutoff point is reduced. Therefore, there is a trade off between validity and the inconclusive rate, which will be seen throughout these findings. One will have to choose between high validity at the cost of a high inconclusive level or a lower level of accuracy in order to reduce the inconclusive rate. In the 7-position scale with a +/-6 cutoff there was a 26% inconclusive rate. However, the accuracy rate for the truthful was found to be 96.5%. A disappointing 89% was found for the deceptive subjects. The total accuracy was 93%. This is certainly inconsistent with the vast majority of studies which have found the reverse: lower rates of accuracy for the truthful subjects and higher degrees of accuracy for the deceptive.

The accuracy rate for the deceptive charts was significantly reduced by four tests in which the two evaluators were in complete agreement with one another, with each reporting high scores, but the original examiners had confirmed all of the results were opposite to what the blind evaluators found. The significance of this will be discussed later in this paper in terms of the reasons for the findings of lower validity in blind scoring research. Excluding inconclusives, the two evaluators were in complete agreement in their decisions, including the four cases in which they were in error. Each evaluator had only the four false negative errors and one evaluator had one false positive error and the other evaluator had two false positive errors.

When the Full DoDPI method was used, the inconclusive rate dropped, but unacceptably high false positive rates were found. Employing cutting scores of +/-4 served to reduce accuracy, but inconclusives as well. Accuracy dropped 3% compared to the +/-6, but inconclusives dropped 12%. Reliability remained almost as high. Using the -6/+4 cutoff has been seen as a means of reducing the false positive rate without affecting the usual high accuracy level with deceptive subjects. The accuracy rate, however, was only 2% less than +/-6, but the difference lay completely in an increase in the false negative rate. Inconclusives were reduced by 8%.

The 3-position scale demonstrated higher levels of both false positive and false negative errors as compared to the 7-position scale with a +/-6 cutoff, but the inconclusive rate was definitely lower. Comparing the +/-3 cutoff with the +/-6 in the 7-position scale the latter was 5.5% more accurate, but the inconclusive rate was 11% greater. Comparing the +/- 4 in the 3-position scale with the +/-6 in the 7-position scale the accuracy for the latter was only 3.5% more and the inconclusive rate was also 3.5% greater. Reliability, which excluded all inconclusives; was somewhat higher for the +/- 4 cutoff, 98% as compared to 93% for the +/-3 and 94% for +/-2. In a comparison between +/-2 in the 3-position scale with +/-6 in the 7-position scale the latter had 5% greater accuracy, but the inconclusive rate was a surprising 18% less.

The +/-2 and +/-3 were essentially equal in accuracy, but +/-4 was approximately

2% greater in the 3-position scale. The inconclusive rate for +/-4 was equivalent to that of +/-6 in the 7-position scale, but the inconclusive rate for +/-3 was only 14.6% and 7.6% for +/-2. In agreement with Krapohl's (1998) findings, which was a laboratory study, and Harwell's (2000) results, that was a field investigation, the +/-4 in the 3-position scale was more similar to the +/- 6 in the 7-position scale than the +/-3 or +/-2 in the 3-position scale. However, the +/-2 might actually be better.

Various indices of reaction were studied including some that had been recommended by DoDPI (Matte, 1996). These findings are shown in Table 2. For respiration, the greatest accuracy was found for suppression 74%, followed by baseline change 71% (rise and fall of baseline), and apnea at the baseline 45%. The few instances of apnea at the ceiling were found to be rare and a chance occurrence. Accuracy at the chance level also was found for a rise in baseline that remained at that level or a drop in baseline that either stayed or returned up again. "Stair steps" up or down were rare and had little significance, as were the occasions of inhalation-exhalation (I/E) ratio change.

Table 2. The validity of certain scoring indices based on 57 confirmed tests in which both evaluators agreed on the findings

Reaction	Number of Occurrences	Accuracy
Suppression	161	75%
Base line change (up & down)	30	71%
Apnea at base of cycle	45	65%
Apnea at ceiling seen as countermeasure	7	57% chance occurrence
Base line loss Drop & remain	25	52% chance occurrence
Base line loss Rise & remain	15	45% chance occurrence
Stair step Excluding Suppression at stimulus After window (relief)	3	33% chance occurrence
Slow respirations Countermeasure	9	55% chance occurrence
Rapid respirations	5	80%
Cardio drop	26	42% chance occurrence

A drop in the cardiograph, which was a fairly frequent occurrence, was found to be at chance level and so was neither an indicator of truthfulness nor deception as has been suggested by DoDPI. There were too few PVC's to evaluate. It had been reported by Barland (1994) that individuals attempting countermeasures sometimes control their breathing and demonstrate slow and regular respirations. Since there were no obvious indications of the use of any countermeasures in these tests, all that could be accomplished in this research in this regard was to evaluate whether deceptive subjects in general showed a slower breathing rate. The respiration rate of deceptive subjects was found to be no slower on the average than those who were truthful, which might only mean that these subjects were simply not using this manner of countermeasure. However, a rapid respiration rate was found much more in some of the deceptive subjects, but they were too few in number to have any significance.

There has been a rather consistent finding in the majority of research that the electrodermal scores provided much of the data for the determination of truth or deception. This was followed by the cardiograph and lastly, respiration. Matte (2000) has disagreed with this placing more value on respiration, but it could be possible that some examiners are simply more proficient in scoring this sensor, and therefore, they obtain more data of value from it. However, in this study, using only those scores obtained from each sensor that were in the accurate direction, it was found that for the truthful subjects the pneumograph contributed 17%, electrodermal 39%, and cardiograph 44%. For deceptive subjects, it was 19%, 50%, and 31%, respectively. The total scores for the truthful and deceptive combined were 17%, 45%, and 37%. Because each of these sensors also contributed results in the direction of inaccuracy, the scores were also calculated to take this into consideration. This time the inaccurate scores were subtracted from the accurate and the final result determined. In this case truthful subjects showed a total of 15% for respiration, 41% for electrodermal, and 44% for the cardiograph, while deceptive subjects were found to have 18%, 60%, and 23%, respectively. The total percentage difference was 16%, 49%, and 34%. The latter approach was seen as being more accurate, but not a very different, representation of the total contribution of the three sensors. Of interest is the fact that deceptive subjects obtained higher scores on the electrodermal sensor and less on the cardiograph. These findings are shown in Table 3.

Table 3. The contribution of each of the three sensors to the total score

	Pneumograph	Electrodermal	Cardiograph
Truthful	15%	41%	44%
Deceptive	18%	60%	23%
Average	16%	49%	34%

Polygraph Validity

This portion of the research will deal with Patrick and Iacono's (1991) notion that a definite difference exists between the scores of confirmed and unconfirmed deceptive polygraph tests. They hypothesized that on the unconfirmed deceptive charts there were likely to be both more truthful subjects and more deceptive individuals who had been found truthful. They stated that none of these subjects would be likely to make any admissions. Therefore, Patrick and Iacono indicated that one could not generalize from the confirmed charts to the unconfirmed. That is, one cannot assume that the same high level of accuracy that exists for confirmed charts can be found in those cases heard in court that have not been confirmed. Because of that, they recommended that polygraph results should not be admitted into evidence.

In this part of the study the scores of 20 randomly selected confirmed deceptive charts were compared to a like number of unconfirmed examinations. The results showed that the total average scores were extremely similar. The confirmed tests averaged -7.6, points and the unconfirmed -8.0 points. These scores tend to be low because the particular police agency where these tests were administered employs the 3-position scale with a +/-3 cutoff. Following this, 24 unconfirmed deceptive tests were tracked through the court system. It was found that in one instance, the case was never heard because the defendant agreed to testify against another individual. Of the remaining 23 cases, 2 were acquitted and the other 21 were found guilty. The percentage of agreement with the unconfirmed polygraph findings was 91%.

The final area of this research relates to previous research that has demonstrated the advantage of scoring to the stronger comparison questions with the first relevant question in a test. Both the studies of Capps and Ansley (1992) and Abrams (1997) have shown that false positives were reduced while there was little risk of causing false negative errors on the findings of deceptive subjects. Since some law enforcement agencies have expanded this approach to include scoring to the stronger comparison questions on two relevants, this study attempted to determine if the latter could result in false negative findings. Of the 20 randomly chosen truthful unconfirmed tests that were re-evaluated by scoring the second relevant to the preceding comparison question, thereby moving all of the scores in the direction of deception, not a single test became deceptive. When 20 additional inconclusive tests were randomly selected with the same procedure employed, 1 of the 20, or 5% of this group, became deceptive with a minimal score of -3. When using the +/-4 cutoff recommended by Krapohl (1998) there were no deceptive findings on any of the 40 tests.

Discussion

Validity

There is little doubt that polygraph validity is very high, certainly higher than most if not all of psychological testing. However, as in any field, some experts are more expert than others and this will have a definite impact on an individual's testing ability and the validity of his or her findings. Moreover, the type of examination and the type of subject will also influence polygraph accuracy. Regardless of who is administering the test it is felt, as Horvath (1992) has indicated, that deceptive subjects are generally more readily detected than truthful subjects are cleared. This has been demonstrated through the consistent findings that indicate that deceptive subjects generally have scores farther from 0, that there is a history of more false positives than false negatives, and that truthful subjects react more to the relevant questions than deceptive individuals react on the comparison questions.

It also has been shown that the original examiners achieve higher rates of accuracy as compared to those who evaluate the tests blindly. This can best be evidenced in Patrick and Iacono's (1987) validity study of original examiners in which they reported 100% accuracy with deceptive subjects and 90% with truthful individuals. These findings are in contrast to their later work where they employed a blind scoring approach and reported 98% accuracy for the deceptive but only 55% for the truthful (1991). Unfortunately, the OTA (1983) findings and Patrick and Iacono's work have been accepted by the American Psychological Association (1986), which has reported that polygraph testing is very inaccurate with the innocent. This was not the case with the American Medical Association (1986) which emphasized employee testing. Patrick and Iacono are still indicating that polygraph validity for the truthful is very low, but it must be recognized that this is an inaccurate assumption. The polygraph charts that they used in their study were administered between 1980 and 1984, which, of course, means that they were using techniques and instrumentation that are now dated. They did not score to the stronger comparison question, and they used the -3 or less spot score rule, both of which result in lower accuracy with truthful subjects. In addition, the present research and Patrick and Iacono's study provide evidence that some examiners utilize extrapolygraphic data in making their decisions. This results in both inaccuracies in blind scoring research and a higher inconclusive rate.

Most research involving large numbers of polygraph tests rely on government or law enforcement charts because of their availability. Some police agencies operate in the same manner as the judicial system in that they make a strong effort not to misjudge an innocent person. Therefore, if the case facts and the subject's behavior suggest that he or she is truthful, even if the subject does not quite reach the cutoff point in numerical scoring, they give the subjects the edge and consider them truthful. The research indicates that they are often accurate in their decision. However, the extrapolygraphic information cannot be found in the tracings so that when blind scoring research is conducted, the accuracy of the truthful has been quite low in some studies. This was demonstrated in 4 of the 10 studies that OTA (1983) chose to evaluate and in Patrick and Iacono's (1991) research where they reported that only 55% accuracy with the truthful was obtained. The original examiner might have made a very conscious decision to use global scoring or it might even have been an unconscious attitude that influenced his or her results. Either way, the original scoring is more accurate and blind scoring is less so.

In the present study, this occurred as well, but with the deceptive subjects. The 96.5% accuracy with the truthful was probably due to the use of better scoring procedures, but the lower rate of accuracy with the deceptive, 89% can only be explained in terms of the four confirmed deceptive charts that both evaluators independently scored as truthful. In fact, after learning that they were in error, they were re-scored but the same results were obtained. The deceptive findings unquestionably were not in the charts, but in spite of that, the original examiners were correct. It is conceivable that they obtained information after the test that indicated that the subject was deceptive or it could have been that the subject was deceptive on the weaker third relevant. Following the DoDPI -3 or less rule, they might have considered the entire test deceptive. What is significant is the effect that the original examiner has on research employing a blind scoring approach. It also demonstrates that Patrick and Iacono (1991) are in error when they state that the only scientific manner of evaluating polygraph validity is through blind scoring in order to eliminate the subjectivity of the original examiner.

Private examiners are generally not able to employ the global approach because their truthful charts typically are evaluated by law enforcement polygraphists who require that all of the data must be found in the tracings. When it is a matter of testing that very likely will be admitted into evidence, the subjectivity that is inevitably a part of decisions made in this manner would not be acceptable in court, despite its greater accuracy. In a courtroom situation, there will almost always be an opposing expert who can only judge the polygraph findings through the tracings; therefore, other influencing factors cannot be employed. This approach might also be unacceptable in a situation in which quality control approaches are employed.

This leaves polygraphy in a dilemma. One cannot effectively employ blind scoring to ascertain the true validity of polygraph testing nor can one utilize the accuracy of the original examiner because of the subjectivity that might exist. This however, is not only a problem in the polygraph field, but one that has existed in psychology and psychiatry, as well as some of the harder sciences. The phrase, "In my professional opinion" is typically employed in court by a professional to indicate that his findings are not only determined by more objective data derived from various testing procedures, but by his or her subjective opinion based on his or her training, experience, and knowledge as well. Perhaps as polygraphy achieves greater acceptance and admissibility, polygraphists too will be accepted to a greater extent and their professional opinions will, as well.

As has been indicated, when one increases accuracy by raising the cutoff points, it also increases the number of inconclusive findings. The 7-position scale using a +/-6 cutoff was found to be more accurate and have the highest reliability, but it also demonstrated the highest inconclusive rate. Therefore, one simply cannot state that one particular position scale with one specific cutoff point is the best possible approach. Perhaps the 7-position scale with a +/-4 is preferable because, while the total accuracy is only 3% less the inconclusive rate is reduced by 12%. However, the false negative rate was quite high. When the -6/+4 was considered, the inconclusive rate compared to +/-6 was 8% less but the accuracy was about 2.5% less. Again, it is all a matter of tradeoff.

None of the 3-position scales was better than the 7-position scale when one considers accuracy and reliability alone. For those who prefer the 3-position scale, the same considerations apply: the closer the cutoffs, the fewer the inconclusives, but there are more errors. For instance, +/-4 produced marginally better accuracy, 89% versus 87% for both +/-2 and +/-3. However, a +/-4 also produced substantially more inconclusive results; 78 % versus 85% for +/-3 and 92% for +/-2. Based on these figures, there is probably cause to consider using +/-2 as cutoffs for the 3-position scale. The +/-4 overall had more inconclusives than +/-3, but generally, +/-4 in the 3-position scale was more similar to +/-6 in the 7-position scale than either +/-3 or +/-2. Comparing the 7-position scale with the 3-position scale, the former had a higher level of accuracy but again slightly more inconclusives.

Since the 7-position scale with a +/- 6 cutoff demonstrated the greatest accuracy and reliability, in spite of the spuriously low validity for the deceptive subjects, for the sake of standardization it is recommended that this particular scoring approach be used.

Indices of Reactivity

The findings related to the indices of reactivity were quite clear. However, it is conceivable that analog instruments would be more likely to demonstrate some other variables that were not seen on the computerized instruments that could be scored. Excluding the amplitude and duration that were seen as valuable indices of reaction on the cardiograph, changes in heart rate were seen infrequently and, therefore, of little value. The drop in cardiograph that had been recommended by DoDPI was found to be a fairly frequent occurrence, but it was not seen as indicative of reaction because it occurred at only chance level. In the electrodermal sensor, amplitude, duration, and complexity were all of value and appeared frequently enough to be significant. Respiration, which is probably the most difficult sensor to score, was found to have value in baseline change, suppression and apnea. Few changes were seen in pattern and rhythm. Regarding scoring in general, the Full DoDPI approach severely impacted on truthful results and it is recommended that this approach not be employed. Other DoDPI recommendations that were of no value were the drop in the pneumograph whether it remained at that level or rose again and the rise in the pneumograph that remained at that level. One might assume that, as Backster has indicated, that this is probably due to a movement of the pneumograph tubes. Capps (2000) indicated that prior to these findings being reported, DoDPI had already eliminated the use of these particular indices of reaction from their teaching program.

Sensors

The evaluation of the degree that each sensor provided information to determine the final score, obtained the same results as prior research. Again the electrodermal sensor was the most valuable, followed by cardiograph, and respiration as a poor third. When a subject is a "pneumo responder" the results can sometimes be very dramatic and helpful in making a decision. It was found that when a subject responded in a specific manner to a particular sensor, it was likely that this would occur several times. If a subject reacted with an apnea, for example, it was not unusual for him or her to continue to react in that manner throughout the examination. This was often true of the other sensors as well. It suggests that while this certainly does not occur in every case, there are individuals who have their own somewhat unique signature within each of the sensors, as well as bodily preferences for specific sensors. The reader is referred to the work of Ansley and Krapohl

(2000) for a more complete study of the frequency of response.

These writers feel that while respiration does not provide as much information as the other sensors, it is a very effective way of monitoring not only countermeasures, but inadvertent deep breaths which might distort the other tracings. Regarding the former, it has been seen that movements and even pain can affect the breathing of the individual. In this way, an evaluation of respiration is a necessity.

Confirmed Versus Unconfirmed Polygraph Charts

Iacono has testified against polygraph admissibility with his strongest arguments being that only blind scoring can measure polygraph accuracy because the original examiners utilize extrapolygraphic information and that most research based on confessions is biased. (*US v Clayton & Dalley*, 1994) Therefore, it was concluded that it is not safe to assume that the research findings obtained on studies that use confession as ground truth can be generalized to those unconfirmed cases heard in court. In the present study, 91% agreement was found between court decisions and unconfirmed deceptive polygraph findings. Because these were unconfirmed decisions on the polygraph, this information never reached either the judge or jury to influence them in any way. Moreover, their notion that higher scores exist on verified tests was not confirmed. Therefore, in direct contrast to Iacono's assumptions, this research indicated that confirmed and unconfirmed findings are of equal value. The high validity found for polygraphy can be assumed to be true of the typical cases admitted into evidence and demonstrates that polygraphy should be considered for admissibility. Moreover, Patrick and Iacono's assumption that only blind scoring is a measure of polygraph validity, was found to be in error. Blind scoring results in more inaccuracies and inconclusives.

Scoring to the Stronger Comparison Question

The present research provided strong evidence that the work of Capps and Ansley (1992) has been very successful in significantly reducing the false positive error rate. Since that time, some law enforcement agencies have expanded this procedure to include scoring to the stronger comparison question on two relevants. However, these writers had no awareness of any research done in this area. Of particular interest is the concern that this could create false negative errors. Truthful and inconclusive unconfirmed tests were employed to ascertain the likelihood of this occurring. Only in one chart of the 20 inconclusives did an individual demonstrate a deceptive chart when the second relevant was scored to the preceding comparison question (-3), and this was confirmed by a court finding of guilty. While additional research should be conducted in this area, these findings suggest that scoring to the stronger comparison question on both relevants will reduce the false positive error to an even greater degree with little risk of creating false negatives. It is recommended that a cutoff of +/-4 be employed with the 3-position scale.

Summary

This study has demonstrated some rather obvious and perhaps some already known polygraph truisms.

1. Polygraph validity is very high, but it is difficult to determine the exact level.

2. The greater the cutoff score the higher the accuracy.

3. The greater the cutoff score the higher the inconclusive rate

4. Conversely, the lower the cutoff score the lower accuracy and the lower the inconclusive rate.

5. The original examiner almost always attains higher accuracy levels.

6. The blind evaluator will obtain lesser accuracy.

7. Since both examiners are evaluating the same charts, it is apparent that the original examiner consciously or unconsciously uses extrapolygraphic data.

8. Polygraphists trained and practiced in the same scoring techniques will obtain the highest reliability.

9. The more that subjectivity is involved in an evaluation, the more likelihood that bias will become an issue. However, these findings suggest that this, in fact, does enhance validity.

10. Polygraphy does not measure truth or deception, only one's physiological reaction to a perceived threat, which the examiner interprets with significant accuracy.

11. Since this accuracy is as great or greater than most psychological tests, polygraph testimony is ready to be considered into evidence.

References

Abrams, S. (1972). Laboratory versus field research. *Polygraph*, 1, 145-150.

Abrams, S. (1987). The heart rate monitor. *Polygraph Update*, 4, 1-2.

Abrams, S. & Davidson, M. (1988). Polygraph countermeasures in polygraph testing. *Polygraph*, 17, 16-20.

Abrams, S. (1997). Stronger versus weaker versus preceding control. *Journal of the American Association of Police Polygraphists*, 1-12.

American Medical Association (1986). Polygraph-Council of Scientific Affairs. *Journal of the American Medical Association*, 256, 1170-1175.

American Polygraph Association (1997). The validity and reliability of polygraph testing. *Polygraph*, 26, 215-239.

American Psychological Association (1986). APA Resolution says reliability of polygraph tests unsatisfactory. News Release, APA, 1200 Seventeenth St. NW Washington, D.C.

Ansley, N & Krapohl, D.J. (2000). The frequency of appearance of evaluative criteria in field polygraph charts. *Polygraph*, 29, 169-176.

Barland, G.H. & Raskin, D.C. (1976). Validity and reliability of polygraph examinations of criminal subjects. Report No. 76-1, Contract No. NI-99-0001 (Washington, D.C. National Institute of Justice, Department of Justice.

Barland, G.H. (1994). Polygraph countermeasures. Presented at the American Polygraph Association Seminar in Nashville, Tennessee.

Berrien, F.K. (1939). A note on laboratory studies of deception. *Journal of Experimental Psychology*, 24, 42-46.

Bradley, M.T. & Janisse, M.P. (1981). Accuracy demonstrations, threat and the detection of deception: cardiovascular, electrodermal, and pupillary measures. *Psychophysiology*, 18, 14 (abstract).

Capps, M.H. & Ansley, N. (1992). Strong control versus weak control. *Polygraph*, 21, 341-348.

Capps, M.H. (2000). American Polygraph Association Seminar, Ft, Lauderdale, Florida.

Dawson, M.E. (1980). Measurements of responses to questions and answers during countermeasure maneuvers. *Psychophysiology*, 17, 8-17.

Harwell, E.M. (2000). A comparison of 3- and 7-position scoring scales with field examinations. *Polygraph*, 29, 195-197.

Honts, C.R. & Hodes, L. (1983). The effects of multiple countermeasures on the detection of deception. *Journal of Applied Psychology*, 47, 405-411.

Honts, C.R., Raskin. D.C., Kircher, J.C. & Hodes, R.L. (1984). Effects of spontaneous countermeasures on the detection of deception. *Psychophysiology*, 20, 583 (abstract).

Horvath, F. S., (1967). The effect of selected variables on interpretation of polygraph records. *Journal of Applied Psychology*, 62, 127-136.

Horvath, F. S. (1972). Personal communication.

Iacono, W.G. (1994). In testimony in *U.S. v. Clayton & Dalley*, No. 92-374-PCT-RCB, Phoenix, Arizona.

Krapohl, D.J. (1998). A comparison of 3- and 7-position scoring scales with laboratory data. *Polygraph*, 27, 210-218.

Kleinmuntz, B. & Szucko, B. (1982). On the fallibility of lie detection. *Law & Society Review*, 17, 84-104.

Kugelmass, S. & Lieblich, I. (1966). Effects of realistic stress and procedural interference in experimental lie detection. *Journal of Applied Psychology*, 50, 211-216.

Lykken, D.T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, 44, 258-262.

Matte, J.A. (1996). Forensic psychophysiology using the polygraph. Williamsville, N.Y., J.A.M. Publications.

Matte, J.A. (2000). Personal communication.

Office of Technology Assessment. (1983). Scientific validity of polygraph testing: a research review and evaluation--a technical memorandum, Washington, D.C. OTA-TM-H-15.

Ohnishi, K., Matsuna, K. & Suzuki, A. (1988). The objective analysis of physiologic indices in the field of detection of deception. *Police Science*, 29, 181-188.

Patrick, C.J. & Iacono, W.G. (1987). Validity of the control question polygraph test: A scientific investigation. *Psychophysiology*, 24, 606 (abstract).

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

Raskin, D.C. (1978). A scientific assessment of the accuracy of detection of deception, *Psychophysiology*, 15, 143-147.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph*, 28, 10-27.

Waid, W.M., Orne, E.C. & Orne, M.T. (1981). Selective memory of social information: alertness and physiological arousal in the detection of deception. *Journal of Applied Psychology*, 66. 224-232.

Short Report
Normative Respiration Data for Criminal Cases

Donald J. Krapohl & Donnie W. Dutton

Key words: field cases, normative data, respiration, Zone Comparison Technique

Respiration is one of three response systems monitored with standard field polygraphs. While much has been written and taught within the polygraph community about phasic responses during deception, tonic respiration during polygraph testing has been given only intermittent attention (Ansley, 1999; Sheve, 1972). Basic issues such as the normal range of respiration frequencies, whether there are gender differences, and whether truthtellers and deceivers have different rates has not been definitively answered. These questions are important for instruction of polygraphy, as well as for the detection of certain types of countermeasures.

To answer these questions, we set about examining respiration characteristics in field examinations. We randomly selected a sample of cases from the confirmed case database of the Department of Defense Polygraph Institute (DoDPI). The sample consisted of 234 first-session criminal polygraph examinations. All examinations were conducted according to the DoDPI Zone Comparison Technique protocol. The cases were collected during a 100% review of all criminal polygraph cases conducted by the US Army Criminal Investigations Division files for a 26-month period beginning on January 1, 1995. Ground truth for all cases had been established independent from polygraph decisions. Each case was recorded on an Axciton computer polygraph (Axciton Systems, Houston, TX). Table 1 shows the composition of the ZCT sample by gender and ground truth.

The goal of this project was to investigate the influence of gender and veracity on tonic respiration rates during criminal polygraph testing. It was expected, based on common assumptions in field practice, that females respire more rapidly than males, and that deceptive examinees will have a higher proportion of slow breathers than non-deceptive examinees.

Table 1. Number of deceptive and nondeceptive examinees, by gender.

	Male	Female	Total
Deceptive	141	33	*174*
Nondeceptive	47	13	*60*
Total	*188*	*46*	*234*

Acknowledgements: The authors are very grateful to Mrs. Kay Williams, whose tireless efforts made this project a success. We are also indebted to the statistical support of Dr. Andrew Dollins and Dr. Stuart Senter of the Department of Defense Polygraph Institute, and Dr. John Stern of Washington University. The opinions expressed in this paper are those of the authors, and do not represent or imply the views of the Department of Defense or the US Government. Requests for reprints should be sent to Donald J. Krapohl, PO Box 10342, Ft. Jackson, SC 29207, or by e-mail to dkrapohl@aol.com.

Results

Average respiration rates for the categories of gender and veracity are found in Table 2. A two-way ANOVA was used to test the effects of the examinee veracity and gender. Veracity did not produce a significant effect on respiration rates (F(1, 233)=0.22, p=0.64). Gender was a significant factor (F(1, 233)=6.02, p=0.02), but there was no interaction for veracity and gender (F(1, 233)=.084, p=0.77).

Table 2. Average respiration rates, with standard deviations, for deceptive and nondeceptive examinees, by gender.

		Male	Female
Deceptive	Average	16.56	18.08
	SD	3.89	3.91
Nondeceptive	Average	16.69	18.62
	SD	3.17	4.25

We constructed a 90% confidence interval around the mean rates for the males and females, rounding values to the nearest whole number. With these data, 5% of examinees would exceed the upper limit, and 5% would fall below the lower limit. See Table 3. Since the veracity of the examinee was not a significant factor, the data were collapsed to gender only.

Table 3. Confidence interval of 90% for respiration rates per minute for males and females, rounded to the nearest whole number.

	Breaths per Minute	
	Lower Limit	Upper Limit
Male	10	23
Female	11	25

Since we did not find differences in tonic respiration between deceptive and nondeceptive examinees, we thought it of interest to investigate changes in respiration rates between individual charts for these groups. A change in rate is a factor worth investigating, since it is generally held in the field that large differences in rate between charts often signals deception. Any such trend could easily be obscured by the present use of averages across charts. A post hoc analysis was conducted to examine rate changes between charts 1 and 2, 2 and 3, and 3 and 1.

The 90% confidence interval for changes in respiration rate between charts for nondeceptive cases was 14.1%. In other words, 5% of nondeceptive examinees slowed their respiration 14.1% or more between charts, and another 5% showed an increase of 14.1% or more. As a concrete example, for an examinee breathing at a rate of 15 cpm on the first chart, a change of 14.1% would be either 12.9 cpm or 17.1 cpm. Taken another way, it is statistically uncommon for a nondeceptive examinee to change his or her tonic respiration rate from 15 cpm to, say, 11 cpm from one chart to another chart. For the deceptive cases, the 90% confidence interval was slightly larger than that of the nondeceptive: 17.5%. Examiners may wish to pay special attention to changes in respiration rates between charts that are unusually large, perhaps greater than 20%. A change of this magnitude between charts does not necessarily signal deception, and should not be considered a decision rule. These are only statistically unlikely behaviors, and warrant an examiner's notice.

Discussion

The finding that tonic respiration rates for deceptive and nondeceptive examinees are not significantly different from one another was unexpected. In the practice of polygraphy, an unusually slow respiration rate (called bradypnea) is often considered a deliberate manipulation by the examinee, and sometimes useful in identifying deceivers. The present data makes clear that the behavior of slow breathing is not unique to either deceivers or truthtellers. As such, in field practice it would be prudent to first determine whether a suspect's breathing rate is genuine or contrived before drawing any conclusions. Even if the tonic breathing rate is being deliberately manipulated by the examinee, that information in isolation is not sufficient to render a decision of deceptiveness or countermeasures. There are other means to make those assessments, and practitioners are directed to the literature relevant to countermeasure detection.

What appears to be meaningful, however, are the very large changes in tonic respiration rates between charts. Because respiration is subject to voluntary control, this pattern is more likely a manifestation of a conscious behavior, not a psychophysiological response, and should be dealt with accordingly. For rounding purposes, changes of about 20% or larger is suggested to trigger these corrective actions, though examiners may choose other thresholds they believe are more appropriate for the conditions of a given examination.

In summary, average breathing rates are different for females and males, but not for deceivers and truthtellers. Very slow or fast tonic respiration rates are not diagnostic in themselves. Examiners should pay attention to examinees who significantly alter the speed of their respiration from chart to chart, especially when the change exceeds 20% between any two charts.

References

Ansley, N. (1999). The frequency of appearance of evaluative criteria in polygraph charts. Final report to Defense Personnel Research Center. ONR Grant Number N00014-98-1-0863

Sheve, W.J. Jr. (1972). Effects of immunizations on polygraph examinations. *Polygraph*, 1(4), 221-233.