THE ACCURACY AND RELIABILITY OF POLICE POLYGRAPHIC ("LIE DETECTOR") EXAMINERS' JUDGMENTS OF TRUTH AND DECEPTION: THE EFFECT OF SELECTED VARIABLES

Dissertation for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

FRANK S. HORVATH

1974

This is to certify that the

thesis entitled

THE ACCURACY AND RELIABILITY OF POLICE POLYGRAPHIC ("LIE DETECTOR") EXAMINERS' JUDGMENTS OF TRUTH AND DECEPTION: THE EFFECT OF SELECTED VARIABLES

presented by

Frank S. Horvath

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Social Science

Victor G. Strecher

Major professor

Date November 14, 1974

O-7639

ABSTRACT

THE ACCURACY AND RELIABILITY OF POLICE POLYGRAPHIC ("LIE DETECTOR") EXAMINERS' JUDGMENTS OF TRUTH AND DECEPTION: THE EFFECT OF SELECTED VARIABLES

By

Frank S. Horvath

The purpose of this study was to determine the accuracy and reliability of judgments of police polygraphic (lie-detector) examiners in blind analysis of polygraphic recordings obtained in field settings; and, to determine whether the accuracy of and confidence in such judgments and the ease with which physiological data were interperted varied according to the particular category from which recordings were drawn and the experience of the examiner.

Method

A stratified random sample of the polygraphic recordings of 112 subjects involved in criminal investigations was drawn from the files of a police agency. Recordings were cross-categorized as verified or unverified, as pertaining to subjects considered truthful or deceptive, and as involving crimes against a person or property crimes.

Ten polygraphic examiners, five with less than three years of experience in lie-detection and five with more, all employed by a law enforcement agency, were recruited to serve as evaluators. Each evaluator independently reviewed the recordings "blind" and indicated: (1) if the subject from whom they were obtained was truthful, deceptive, or inconclusive; (2) his degree of confidence in each truth/deception judgment; and (3) the ease of interpretability of each of three physiological indices, respiratory, electrodermal (GSR), and cardiovascular.

Analysis

Hypothesis-testing procedures were carried out using analysis of variance in a 2.2 x 2 x 2 Split-plot design. The four factors were: Experience (high/low); Verification (verified/unverified); Truthfulness (truthful/deceptive); Crime-type (person/property). Dependent variables treated separately were accuracy scores, the percentage of correct judgments; confidence scores, the sum of confidence ratings; and total ease-of-interpretability scores, the sum of the "ease" ratings for the three physiological indices.

Results

Overall, the evaluators made 63.1% correct judgments ($p < .001$). Contrary to expectations, high-experience evaluators were neither more accurate ($p > .10$) nor confident ($p > .10$) in their judgments nor did they consider recordings easier to interpret than did low experience evaluators ($p > .10$).

Predicted main-effects for the Verification, Truthfulness, and Crime-type factors for all three dependent variables were complicated by interactions. In essence, analysis of these interactions indicated that: recordings in the "deceptive/crime against a person" categories were judged more accurately, and those in the "truthful/crime against a person" categories less accurately, than all others across levels of verification; and that recordings of deceptive subjects were judged with greater confidence and were easier to interpret than those of truthful subjects irrespective of the nature of verification.

Intra-class correlation-coefficients calculated separately for evaluators' judgments of verified and unverified recordings indicated that the judgments in both of these conditions were highly reliable, .89 and .85, respectively.

Both confidence-ratings and total "ease" ratings were higher in correct than in incorrect judgments ($p < .002$; $p < .001$, respectively). Further analysis of the ease-of-interpretability ratings indicated that evaluators rated respiration, cardiovascular activity, and GSR easier to interpret, in order; ratings were higher in correct than in incorrect judgments for respiration ($p < .001$) and cardiovascular activity ($p < .001$), but not for GSR ($p > .10$).

Other issues investigated showed that accuracy increased as the number of evaluators in agreement increased, and that accuracy was higher ($p < .001$) when evaluators'

judgments were based on recordings with less rather than more polygraphic data. The results of a numerical scoring-scheme, as carried out by evaluators on a sub-sample of recordings, indicated that GSR-scores were more accurate than were those of the other two indices if inconclusive scores were eliminated, and that GSR was scored more consistently than either respiration or cardiovascular activity.

Methodological differences between this and other research on the same topic are presented to account for some of the differences in results. Further, it is suggested that differences between polygraphic recordings, due to the nature of lie-detection in the field, account for some of the observed interaction effects.

THE ACCURACY AND RELIABILITY OF POLICE POLYGRAPHIC ("LIE DETECTOR") EXAMINERS' JUDGMENTS OF TRUTH AND DECEPTION: THE EFFECT OF SELECTED VARIABLES

By

Frank S. Horvath

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

College of Social Science

1974

Dedicated to

Jan and Juliann

ACKNOWLEDGMENTS

The writer was requested not to identify either the police agency or the polygraphic examiners who provided for and took park in this study. In spite of their anonymity, however, the writer gratefully acknowledges their interest, cooperation, and assistance; a sincere thanks is extended to all concerned.

In addition the writer is indebted to the following persons and organizations: Dr. Steven Olejnik, Office of Research Consultation, Michigan State University; Mr. James Mullin, Office of Applications Programming, Michigan State University; and Dr. Charles Hanley, Assistant Dean for Graduate Education, College of Social Science, Michigan State University, for their advice and assistance concerning the statistical treatment and evaluation of the data; Dr. Victor G. Strecher, Dr. Robert C. Trojanowicz, Professor Ralph F. Turner, Dr. Lawrence I. O'Kelly, Dr. Hiram Fitzgerald, and Dr. Peter K. Manning, members of my Ph.D. committee, for their interest and advice; and to the U.S. Department of Justice, for its financial support through the L.E.A.A. Graduate Research Fellowship Program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure

4.10--M
d
a
v

Figure Page

LIST OF APPENDICES

lying

sure,

For t

the a

what

assoc

anxie

patte

quest

discr

lying

data

"lie

Effec

Gesa

Inst

J. B

1-33

With

Appl

Bloo

117-

Doub

CHAPTER I

INTRODUCTION

It has long been known that under certain conditions lying is accompanied by changes in heart rate, blood pressure, breathing, and electrical conductivity of the skin.[1] For the most part these responses are under the control of the autonomic nervous system, and to a lesser degree, somewhat under voluntary control. Although the responses associated with lying are also characteristic of arousal, anxiety, stress, etc., it is possible that discernable patterns of physiological response to appropriately framed questions within a structured setting do make possible discrimination between persons telling the truth and persons lying. Such discrimination based upon recorded physiological data forms the basis for the procedure popularly known as "lie detection."

---

[1] V. Benussi, "Die Atmungssyptome der Lüge" ("On the Effects of Lying on Changes in Respiration"), Arch. für Die Gesamte Psychologie, 31 (1914), 244-273; H. Burtt, "The Inspiration-Expiration Ratio During Truth and Falsehood," J. Exp. Psych., 4 (1921), 1-23; N. Chappell and N. Matthew, "Blood Pressure Changes in Deception," Arch. Psych., 17 (1929), 1-39; F. Peterson and C. Jung, "Psychophysical Investigations with the Galvanometer and Pneumograph in Normal and Insane Individuals, Brain, 30 (1907), 153-218; W. Marston, "Systolic Blood Pressure Symptoms of Deception," J. Exp. Psych., 2 (1917), 117-163; H. Munsterberg, On the Witness Stand (New York: Doubleday, 1908), 118-133.

been u

for at

recent

who pr

purpos

of the

basis

tion.

partic

work.4

bility

by the

Actually, lie detection in one form or another has been used to determine the truthfulness of criminal suspects for at least the past fifty years.[2] And, while within recent years there has been a marked proliferation of persons who practice it for both law enforcement and commercial purposes,[3] surprisingly little is known about the validity of the procedure or the reliability of decisions made on the basis of it.

There are several reasons for this lack of information. First, polygraph examiners themselves have not been particularly prone to offer proof of the efficacy of their work.[4] Second, research concerning the validity and reliability of real-life (field) lie detection has been hampered by the lack of an acceptable "ground truth" criterion for

---

[1]P. Trovillo, "A History of Lie Detection," J. Crim. Law, Crim., 29 (1939), 848-881 and 30 (1939), 104-119; W. Marston, The Lie Detector Test (New York: Richard K. Smith, 1938); J. Larson, Lying and Its Detection (Chicago: Univ. of Chicago Press, 1932. Reprinted, Montclair, N.J.: Patterson Smith, 1969).

[3]See: N. Ansley (Ed.), "Actions of the Board of Directors, January 18-20," American Polygraph Association Newsletter, No. 1 (Dec.-Jan., 1974), 10; R. Paterson, "The Future of Polygraph in Industrial Security," American Polygraph Association Newsletter, No. 8 (Sept. 1972), 1-3.

[4]Much of this research reported by examiners has been criticized on methodological and other grounds. See: R. Sternbach, L. Gustafson and R. Colier, "Don't Trust the Lie Detector," Harv. Bus. Rev., 40 (1962), 127-134; J. Orlansky, An Assessment of Lie Detection Capability (Declassified Version), Tech. Rep. 62-16 (Arlington, Va.: Inst. for Defense Analyses, Res. and Eng. Support Div., July, 1964, 6-18.

validity-studies, and the lack of standardized testing procedures for reliability studies.[5] Third, the bulk of research in lie-detection has been done in the laboratory where adequate control over data-collection, ground-truth criteria, etc. is possible, but where results do not necessarily pertain to conditions outside the laboratory.[6] Finally, field lie-detection is essentially an empirically developed procedure with a minimal theoretical foundation; it is an art, not a science.

Within recent years research in lie-detection has received considerable attention from field practitioners and, within the scientific community, psychologists and psychophysiologists. The major thrust of field research has been toward validation and improvement of current practices; that of scientific research has been to uncover the precise physiological, and particularly psychological, mechanisms which make lie-detection feasible. In spite of this split in direction of research, there is wide agreement that lie-detection works.[7] Exactly how well it works in the field, how valid and how reliable its indications of truth and deception are, these are questions provocative of a

---

[5] M. Orne, R. Thackray and D. Paskewitz, "On the Detection of Deception: A Model for the Study of the Physiological Effects of Psychological Stimuli," Handbook of Psychophysiology, N. Greenfield and R. Sternbach (Eds.) (New York: Holt, Rinehart and Winston, 1972), 743-785.

[6] M. Orne, "Implications of Laboratory Research for the Detection of Deception," Polygraph, 2 (1973), 169-199.

[7] Ibid., 177.

healthy

Can the

reflec

two ex

interp

same wa

the "ac

polygra

of poly

blind a

on the

sidere

underg

consid

terist

person

common

bility

sugges

daily

Lie De
44 (19
to Tru
Sci. a

healthy skepticism among field examiners and researchers. Can the judgment of a polygraphic examiner be an accurate reflection of a person's truthfulness or deception? And will two examiners, or the same examiner at two different times, interpret the same set of polygraphic recordings in the same way?

Purpose of The Study

The primary purpose of this study was to determine the "accuracy" and reliability of judgments made by trained polygraphic examiners; the technique used was blind analysis of polygraphic recordings obtained in field settings. In blind analysis judgments of truth telling and lying are made on the basis of polygraphic records exclusively; not considered are such aspects as behavioral cues of a person undergoing examination, investigators' reports and opinions, consideration of age, sex, race and other personal characteristics, or the examiner's intuitive response to the person being examined. Such sources of information are commonly believed to contribute to the validity and reliability of lie-detection.[8] However, as recent research suggests[9], current testing procedures which include individually distinct response patterns to control questions, make

---

[8]See: J. Reid and R. Arther, "Behavior Symptoms of Lie Detector Subjects," J. Crim. Law, Crim., and Pol. Sci., 44 (1953), 104-108; F. Horvath, "Verbal and Nonverbal Clues to Truth and Deception During Polygraph Examinations," J. Pol. Sci. and Adm., 1 (1973), 138-152.

[9]This research discussed in detail in the next chapter.

lie-d

infor

to st

exami

evalu

in su

istic

it de

exari

The c

was r

his d

zatic

of po

These

detec

attai

purpo

limit

simil

Inter
V. Le

graph
Law,
P. As
Diagn

lie-detection relatively independent of outside sources of information. That is, control-question testing is believed to standardize lie-detection so that judgments made by an examiner in actual testing and by trained, independent evaluators of the polygraphic recordings thus obtained, are in substantial agreement.

This study incorporated several design characteristics which distinguish it from previous research. First, it dealt exclusively with judgments made by polygraphic examiners (evaluators) employed by law-enforcement agencies. The only prior research having some bearing on this issue was reported by Holmes, who, unfortunately, did not report his data in sufficient detail to allow for valid generalizations.[10] Other research was concerned with the judgments of polygraphic examiners employed by a commercial agency. These examiners received more initial training in lie-detection theory and practice and had higher educational attainment than most examiners employed for law enforcement purposes;[11] it is likely that their training and education limit generalization of their results to examiners having similar backgrounds.

---

[10] W. Holmes, "The Degree of Objectivity in Chart Interpretation," Academy Lectures on Lie Detection, Vol. II, V. Leonard (Ed.) (Springfield, Ill.: C.C Thomas, 1958), 62-70.

[11] F. Horvath and J. Reid, "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," J. Crim. Law, Crim. and Pol. Sci., 63 (1971), 276-281; F. Hunter and P. Ash, "The Accuracy and Consistency of Polygraph Examiner's Diagnoses," J. Pol. Sci. and Adm., 1 (1973), 370-375.

graphi

invest

only f

fessio

ments

criter

simila

for th

agreem

defini

does n

validi

estima

themse

may co

sugges

(prior

ent fr

simila

in inv

fessio

so ver

etc. n

the De

Second, judgments were made by evaluators of polygraphic recordings drawn from both verified and unverified investigations. Previous research utilized recordings drawn only from verified investigations, using corroborated confessions as the criteria of verification; accuracy of judgments was then assessed in terms of agreement with the criteria. In this study, however, while accuracy was similarly defined for judgments made on verified recordings, for those made on unverified recordings it was defined as agreement with the testing examiner's judgment. Such a definition, of course, has serious disadvantages since it does not allow for any conclusions to be drawn about the validity of judgments, but it is a useful definition for estimating the contribution which the polygraphic recordings themselves make to lie detection.

It is clear that the use of only verified records may considerably bias research. For instance, it has been suggested that persons presumed by examiners to be liars (prior to testing) may undergo examinations somewhat different from those presumed to be truth-tellers.[12] Using similar reasoning one could conclude that persons involved in investigations which are eventually "verified" by confession might undergo examinations differing from those not so verified; factual information, behavioral characteristics, etc. might provide more, or "better" clues in the verified

---

[12]Orne, "Implications of Laboratory Research for the Detection of Deception," op. cit., 176.

invest

for so

of ind

ting t

record

only a

verifi

verifi

unveri

record

study.

concer

It is

crimes

factua

robber

detail

suspe

for f

Inter
113.

the p
volve
and i
"crim
break
destr
volve

investigations; or, perhaps, the resulting polygraphic records, for some reason, might be of a better quality to the advantage of independent evaluation. Furthermore, the need for evaluating the judgments made of both verified and unverified records is readily apparent when one considers the fact that only a small proportion of all polygraphic examinations are verified by any means at all.[13] Findings based only upon verified records are not necessarily applicable to the unverified situations.

Third, the nature of the investigation from which recordings were drawn was incorporated in the design of the study. That is, recordings were drawn from investigations concerning crimes against a person and property crimes.[14] It is apparent when considering these two categories of crimes that an examiner usually has access to more detailed factual information in the former; a victim of an armed robbery, for instance, is usually capable of relating precise details of the crime, and in some cases, of identifying a suspect. Such detailed information provides a firmer basis for formulating appropriate test questions which, as field

---

[13] F. Inbau and J. Reid, Lie Detection and Criminal Interrogation (Baltimore: Williams and Wilkins, 1953), 110-113.

[14] The criterion used for classification of crimes was the presumed nature of involvement of the victim; direct involvement, such as in rape, murder, armed robbery, assault, and indecent (sexual) liberties, led to classification as "crimes against a person." On the other hand, crimes such as breaking and entering (burglary) arson, larceny, malicious destruction of property, and embezzlement, when victim involvement is less apparent, were classified "property crimes".

examine

physiol

Consequ

investi

accurat

ability

experie

if expe

than th

that ev

polygra

accurat

who had

differe

that th

group.

in a m

evalua

those

Polygr

Wilkin

J. Pol

Examin

examiners are well aware, are important determinants of physiological responsiveness during polygraphic examinations.[15] Consequently, it can be suggested that recordings drawn from investigations involving crimes against a person may be more accurately judged than those involving property crimes.

Fourth, although previous research suggests that the ability to interpret polygraphic records is a function of experience, it is not clear if such a finding would pertain if experience were defined in a manner somewhat different than that reported. For instance, Horvath and Reid found that evaluators with less than six months of experience (in polygraph testing), and still undergoing training, were less accurate and consistent in their judgments than evaluators who had completed their training.[16] Certainly, such a difference is reasonable since one would not anticipate that the untrained evaluators would do as well as the other group. Hence, in this study experience levels were defined in a more meaningful manner, although it was anticipated that evaluators with more experience would be more accurate than those with less experience.

---

[15]See: J. Reid and F. Inbau, Truth and Deception, The Polygraph ("Lie Detector") Technique (Baltimore: Williams and Wilkins, 1966), 1621; R. Arther, "Crime Question Wording," J. Polygraph Studies, 4 (Sept.-Oct., 1969), 1-4.

[16]Horvath and Reid, "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," op. cit., 278-279.

Fifth, the recordings used in this study constituted a random sample of a pre-defined population. This is in contrast to previous research dealing only with recordings chosen in accordance with some arbitrary criterion and which, moveover, substantially controlled the nature of the interaction between the examiner and examinee (subject). For instance, Horvath and Reid reported results obtained when evaluators judged recordings selected because they were believed to require sufficient skill to interpret. Moreover, the recordings used by Horvath and Reid were obtained from subjects who were tested by only one examiner. The use of such recordings at least partially controls for the nature of the interaction between the examiner and subject, interaction believed to have an affect on the nature of the recordings obtained.[17]

It is not known if, when such interaction is not controlled, judgments of independent evaluators would be accurate and in substantial agreement with the testing examiner. But it is clear that proponents of control-question testing maintain that such would be the case.[18]

A second purpose of this study was to employ several devices used in experimental "lie detection" studies but

---

[17]Orne, "Implications of Laboratory Research for the Detection of Deception." op. cit., 175-177.

[18]Horvath and Reid, "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," op. cit., 281; Hunter and Ash, "The Accuracy and Consistency of Polygraph Examiner's Diagnosis," op. cit., 375.

not u

ings

the d

evalu

of th

lie-d

of re

the r

exami

with

to th

repor

confi

than

in th

ratir

inexp

parti

bilit

Armed

prepa

(602)

of Pe

Univ.

7125)

bili-

not used at all in studies dealing with polygraphic recordings obtained from field settings. First, evaluators rated the degree of confidence in their judgments. Second, evaluators indicated the "ease-of-interpretability" of each of the three basic physiological measures used in field lie-detection. And, finally, evaluators judged a sub-sample of recordings in accordance with a numerical scoring system, the reliability of which, although developed by a field examiner, has not been reported in the literature dealing with evaluations of field-derived recordings.

The confidence scale used in this study was similar to that employed by Kubis[19] and Moroney,[20] both of whom reported similar results: independent evaluators had "greater confidence in those decisions ultimately verified as correct than they did in those which were incorrect."[21] The scale in the present study was used to determine if confidence ratings were higher for experienced evaluators than for inexperienced; if such ratings varied depending upon the particular category from which polygraph recordings were

---

[19] J. Kubis, Studies in Lie Detection: Computer Feasibility Considerations, Tech. Report 62-205 (Arlington, Va.: Armed Services Technical Information Agency, June, 1962), prepared for Air Force Systems Command, Contract No. AF 30 (602)-22700, Project No. SS34, Fordham University, 1962, 146.

[20] W. Moroney, "The Detection of Deception as a Function of PGR Methodology" (unpublished Ph.D. dissertation, St. John's Univ., 1968, Ann Arbor, Mich.: Univ. Microfilms, 1969, No. 69-7125).

[21] Kubis, Studies in Lie Detection: Computer Feasibility Considerations, op. cit., 68.

drawn

would

inter

ilar t

deter

easie

tabil

from

that

to in

(GSR)

or ca

such

obta

seem

of t

valu

vasc

to i

grap

to w

prim

sure

brea

elec

meas

drawn; and, if, as Kubis and Moroney found, greater confidence would be indicated in correct than in incorrect judgments.

The scale used in this study dealing with the "ease-of interpretability" of the various physiological measures was similar to that reported by Kubis.[22] The purpose of the scale was to determine if more experienced evaluators judged recordings easier to interpret than less experienced; if ease-of-interpretability ratings varied depending upon the particular category from which recordings were drawn; and, if, as Kubis found, that records on which correct judgments were made were easier to interpret than those judged incorrectly.[23]

Kubis reported that the psychogalvanic response (GSR) was judged easier to interpret than either respiratory or cardiovascular measures. It is difficult to predict that such a result would pertain in evaluations of recordings obtained from field settings, although such an expectation seems reasonable, primarily because of the simple wave-form of the GSR. However, most field-examiners disclaim the value of GSR and give precedence to respiratory and cardiovascular activity;[24] hence, it is possible that either of

---

[22] Ibid., 146. [23] Ibid., 71.

[24] Throughout this paper terms of convenience are used to identify the physiological parameters recorded by the polygraph instrument. Cardiovascular activity or "cardio" refers to what is commonly termed the "blood-pressure-pulse rate," primarily a measure of complex interaction between blood pressure and volumetric changes; respiration refers to changes in breathing rate and volume; galvanic skin response (GSR) and electrodermal activity are used interchangably, typically measures of the Skin Resistance Response.

the t

pret

tatio

polyg

ing s

only

in an

obtai

this

numer

vario

by tr

of li

studi

Evalu
Assoc

bili

the two latter measures would be judged easier to interpret than GSR because of the particular training and orientation of field examiners.

Evaluators in this study analyzed a sub-sample of polygraphic recordings in accordance with a numerical scoring system, the reliability of which has been reported in only one study.[25] Such a system, however, has not been used in any reported study dealing with polygraphic recordings obtained from field settings. Hence, it was of interest in this research to explore the overall reliability of the numerical scoring system and to determine which of the various physiological measures was most reliably evaluated by trained field-examiners.

Need for the Study

Orlansky, in his assessment of the state of the "art" of lie-detection reported that:

> Except for Kubis (1962) no one has explored the possibility that two examiners working independently might make different interpretations of the same record. Reliability of the polygraph in the sense of consistency of measurement, i.e., agreement among examiners, is an unknown quantity.[26]

Since Orlansky's report there have been only two studies conducted to determine the accuracy and reliability

---

[25] G. Barland, "The Reliability of Polygraph Chart Evaluations" (paper presented at The American Polygraph Association Seminar, August 15, 1972, Chicago, Ill.).

[26] Orlansky, An Assessment of Lie Detection Capability, op. cit., 8.

of

set

invo

rou

Both

exa

publ

stu

tra

dete

high

ques

dete

Syst

fede

admi

evid

essa

and

it h

cour

tanc

disc

adm
10:3

of "blind" judgments made on data obtained from field settings. Although there have been other such studies involving experimental lie detection, none of them can be routinely generalized as pertinent to the field situation. Both of the field studies were based on judgments made by examiners trained in the same manner and not employed by a public law enforcement agency. Hence, in spite of these studies we still do not know if polygraphic examiners, trained in a somewhat different manner and engaged in lie-detection specifically for police purposes, can achieve high reliability in their decisions. An answer to this question would not only extend our knowledge about lie-detection, but bare implications for our Criminal Justice System, particularly the courts, as well.

During the past fifty years only one of the reported federal and state court decisions considering the question admitted unstipulated polygraphic examination results as evidence. The reasons for this exclusionary policy were essentially that the polygraphic technique lacked reliability and a "general acceptance" in the particular field in which it belongs.[27] Recently, however, there have been several court decisions indicating a trend to wider judicial acceptance of the technique. Altarescu has published an excellent discussion of these decisions and the problems remaining for

---

[27]The "general acceptance" test concerning polygraph admissibility was set out in Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).

the p

Predi

of th

in ex

emplo

groun

techn

over,

graph

shoul

previ

all

revi

the polygraphic field itself if the trend continues.[28] Predictably, one of these problems concerns the reliability of the technique, especially in regard to examiners who vary in experience, qualifications, and particular technique employed.

It is hoped that this study will provide a firmer ground for answers to questions concerning the polygraphic technique which have for so long troubled our courts. Moreover, as the study deals directly with reliability of polygraphic examiners employed by police agencies, the results should have a more direct impact on the judiciary than previous studies.

---

[28] H. Altarescu, "Problems Remaining for the 'Generally Accepted' Polygraph," reprinted from: Boston Univ. Law Review, 53 (March, 1973), 375-405.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

Essentially the literature dealing with lie-detection can be identified as that written by field practitioners and that written by laboratory researchers. Literature in the former category usually consists of descriptions of procedures, instrumentation and some research bearing on the efficacy of these items. On the other hand, reports of laboratory researchers most often are concerned with determining how well and under what conditions lie-detection is possible; that is, what precise physiological and psychological mechanisms contribute most to the detection of deception. Because both goals and methods of these two approaches differ, the literatures will be dealt with separately, considering first procedural differences. The relatively detailed discussion of field procedures will not only provide a more thorough base for assessment of laboratory procedures, but will also clarify points to be made in discussion of the validity and reliability of lie-detection. But, first, a historical review of lie-detection is in order.

of l

grap

acco

foll

chap

date

tong

burr

are

vati

in p

abou

phys

nogr

ment

dete

took

In l

Crim
1/97
Pres
1963
fiel

Historical Evaluation

There is no need to discuss in depth the early history of lie-detection procedures and the development of the polygraph instrument, as there are already available excellent accounts dealing with this topic.[1] The purpose of the following brief review of this area is simply to put this chapter into perspective.

Historically, the most dramatic attempts at lie-detection relied upon "ordeals" such as hot irons on the tongue of suspects to be protected by their innocence or burned by their guilt. Also described in the literature are relatively objective procedures, such as careful observation of a suspect's behavioral characteristics or changes in pulse rate when under interrogation. It was not until about 1895, however, when Cesare Lombroso, an Italian physiologist, and his student, Mosso, used the hydrosphygmograph and the "scientific cradle", that objective measurement of physiological changes became associated with the detection of deception.[2]

Following Lombroso and Mosso, other investigators took note of physiological changes associated with deception. In 1908 Munsterberg made reference to the effect of lying on

---

[1]See: P. Trovillo, "A History of Lie Detection," J. Crim. Law and Crim., 29 (1939), 848-881 and 30 (1939), 104-119; J. Larson, Lying and Its Detection (Chicago: Univ. Chicago Press, 1932, reprinted, Montclair, N.J.: Patterson Smith, 1969); C. Lee, The Instrumental Detection of Deception (Springfield, Ill.: C.C Thomas, 1953).

[2]Trovillo, "A History of Lie Detection," op. cit., 858.

brea

and

cond

tion

ing

Burt

indi

find

syst

or G

bloc

dure

puls

deve

used

atus

acti

Doub

Effe
Test
a H

Trut
also
Rati

Deco

Test

J. P

breathing, cardiovascular activity, involuntary movements, and the galvanic skin response (GSR).[3] In 1914, Benussi conducted a series of experiments in which he found a relationship between the inspiration-expiration ratio in breathing and deception.[4] His findings were later confirmed by Burtt who added that systolic blood pressure was yet more indicative of deception than respiration.[5] Marston's findings agreed with Burtt's that discontinuous measures of systolic blood pressure were superior to either respiration or GSR for detecting deception.[6] Larson modified Marston's blood pressure test and developed an instrument and procedure for making continuous recordings of both blood pressure-pulse rate and respiration.[7] Keeler, generally credited with developing the prototype of the polygraph instrument now used in most field settings, further refined Larson's apparatus to which he added a device for measuring electrodermal activity.[8]

---

[3] H. Munsterberg, On The Witness Stand (New York: Doubleday, 1908), 118-133.

[4] V. Benussi, "Die Atmungssymptome der Lüge" ("On The Effects of Lying on Changes in Respiration"), Arch. für Die Gestamte Psychologie, 31 (1914), 244-273, cited by Trovillo, "A History of Lie Detection," op. cit., 870.

[5] H. Burtt, "The Inspiration-Expiration Ratio During Truth and Falsehood," J. Exp. Psych., 4 (1921), 1-23; see also, H. Burtt, "Further Technique For Inspiration-Expiration Ratios," J. Exp. Psych., 4 (1921), 106-110.

[6] W. Marston, "Systolic Blood Pressure Symptoms of Deception," J. Exp. Psych., 2 (1917), 117-163.

[7] J. Larson, "Modification of The Marston Deception Test," J. Amer. Inst. Crim. Law and Crim., 12 (1921), 390-399.

[8] L. Keeler, "A Method For Detecting Deception," Amer. J. Pol. Sci., 1 (1930), 38-52.

The discussion up to this point should not be taken as an indication that respiration, cardiovascular activity, and GSR are the only physiological processes which have been associated with deception. Limited success at detecting deception has also been accomplished by measurement of other physiological activity, such as: hand tremors,[9] electroencephalic activity,[10] pupil dilation,[11] oculomotor activity,[12] voice modulation,[13] oxygenation of the vascular system,[14] and

---

[9]A. Luria, "The Union of the Motor Method and the Investigation of the Affective Reaction," State Inst. of Exp. Psych. (Moscos, 1928); "Die Methode der Abbildenden Motorik und ihre Anwendung an die Affekt-Psychologie, Psychol-Forschung, Band 12, 1929; Examination and Psychical Reactions (1930); The Nature of Human Conflicts, Horsley Gannt (Trans. and Ed.), 1932, cited by Trovillo, "A History of Lie Detection," op. cit., 114, note 124.

[10]C. Oberman, "The Effect on the Berger Rhythm of Mild Affective States," J. Abn. and Soc. Psych., 34 (1939), 84-95.

[11]F. Berrien and G. Huntington, "An Exploratory Study of Pupillary Responses During Deception," J. Exp. Psych., 32 (1943), 443-449.

[12]F. Berrien, "Ocular Stability in Deception," J. App. Psych., 26 (1942), 55-63; F. Berrien, "Possibilities in The Use of The Opthalmograph as a Supplement to Existing Indices of Deception," Psych. Bulletin, 37 (1940), 507; D. Ellson, R. Davis, I. Saltzman and C. Burke, A Report of Research on Detection of Deception (Tech. Report prepared for Office of Naval Research, Contract N6onr-18011, Indiana Univ., 1952).

[13]M. Alpert, R. Kurtzberg, and A. Friedhoff, "Transient Voice Changes Associated with Emotional Stimuli," Arch. Gen. Psych., 8 (1963), 362-365; P. Fay and W. Middleton, "The Ability to Judge Truth-Telling or Lying From the Voice Transmitted over a Public Address System," J. Gen. Psych., 24 (1941), 211-215.

[14]H. Dana, "It is Time to Improve the Polygraph: A Progress Report on Polygraph Research and Development," Academy Lectures on Lie Detection, II, V. Leonard (Ed.), (Springfield, Ill.: C.C Thomas, 1957), 84-90; H. Dana and C. Barnett, "The Emotional Stress Meter," Academy Lectures on

cover

upon b

tion

vascu

some

instr

thou

comm

in u

ques

thes

inte

rel

covert muscular movements.[15] But what is now fairly well agreed upon by field examiners is that any attempt at detecting deception must be made with an instrument that records both cardiovascular and respiratory activity.[16] It is in fact illegal in some states for a "detection of deception" examiner to use an instrument not capable of recording these two parameters, although others, particularly electrodermal activity are also commonly recorded in conjunction with them.[17]

Field Lie Detection: Procedures

There are two major field lie-detection procedures in use today, the relevant-irrelevant (R-I) and the control-question (CQ) techniques. In this section a discussion of these techniques will be made in some detail, to aid in an understanding of the literature concerning the validity and reliability of lie-detection.

Relevant-Irrelevant Technique

It is clear from the literature on field lie-detection that many of the early practitioners considered the primary

---

Lie Detection (Springfield, Ill.: C.C Thomas, 1957), 73-83; R. Thackray and M. Orne, "A Comparison of Physiological Indices in Detection of Deception," Psychophysiology, 4 (1968), 329-339.

[15] J. Reid, "Simulated Blood Pressure Responses in Lie Detector Tests and a Method for Their Detection," J. Crim. Law and Crim., 36 (1945), 201-214.

[16] N. Ansley (Ed.), "Inquiry Regarding Dektor PSE-1," American Polygraph Association Newsletter, Number 3 (March, 1972), 18.

[17] C. Romig, "The Status of Polygraph Legislation of the Fifty States," Part III, Police, 16 (1971), 58.

benefit of polygraphic testing to be that it enhanced their own ability to obtain confessions of guilt or admissions of lying from criminal suspects.[18] It is not surprising then that polygraphic testing and "interrogation" (intensive or accusatory questioning designed to secure a confession) were often considered identical, and perhaps inseparable, processes; that is, the two processes were blended or combined in such a way that the psychological effect of the polygraphic instrument and the consequent physiological recordings could be maximized to secure confessions of guilt. The complete blending of interrogation and polygraphic testing characterizes the R-I technique.[19]

Pre-Test interview.--Simply stated, the R-I Technique is relatively unstructured, consisting of an interview, or perhaps intensive questioning, followed by or combined with polygraphic testing. During the interview the examiner discusses with the subject background information relative to the investigation at hand and exploits any hesitancy or uncertainty in the subject's answers to questions, he also observes the

---

[18]See: F. Inbau, Lie Detection and Criminal Interrogation (Baltimore: Williams and Wilkins, 1942), 54.

[19]The R-I Technique is considered outmoded by some leading examiners: See: C. Backster, "Lie Detection Comes of Age," Law and Order (undated, unpaginated reprint supplied by author); C. Backster, "Methods of Strengthening our Polygraph Technique," Police, 6 (1962), 61-68.

subject's behavior in order to locate "sensitive areas" which may be useful in the testing. The examiner also explains the purpose of the testing and the nature of the polygraphic instrument, implying that it is futile for the subject to harbor any thoughts of "beating" the test. It is also the examiner's purpose during the interview to establish rapport with the subject and to become familiar with his language and personal history in order to assure that the test questions, which may or may not be reviewed prior to testing, will be effectively worded.

The length of the interview is determined by the examiner according to his impression of the subject's emotional accessibility. A high-strung subject generally requires a lengthier interview in order to prepare him for testing; a relatively passive subject must be "aroused", and so forth.

Polygraphic testing.--Polygraphic testing in the R-I Technique generally consists of asking a series of questions relevant to the crime and interspersed between irrelevant, or non-critical questions; other types of questions such as those exposing a guilt complex may be asked at the discretion of the examiner. The precise nature, wording, and ordering of the test-questions is determined by the examiner as testing progresses, as is the length of any one test. Generally, however, generalized questions precede specific questions, an order believed helpful because it recapitulates the steps in commission of an offense.

re

is

up

ap

in

co

is

ca

fu

ir

wo

up

th

du

en

of

be

co

pr

ex

ac

qu

th

of

The length of any given test, the asking of the relevant and irrelevant questions at least once in a series, is determined by the examiner and is dependent primarily upon the subject's ability to withstand the effects of the apparatus used for recording cardiovascular activity. Within any given polygraphic examination, two R-I tests may be conducted before a determination of deception (or truthfulness) is made, although proponents of the method feel that in most cases such a determination can be made following one test.

Proponents of the R-I technique assume that truthful people will not differentially react to relevant and irrelevant questions, while people lying will. In other words, determinations of truth-telling and lying depend upon perceptible differences in physiological response to the stimulus of non-critical and critical items. Moreover, during any given test or between any two tests such differential reactions constitute cause for intensive questioning of the subject by the examiner. Proponents of this technique believe that "interrogation" for the purpose of securing a confession or admission of lying at any time during the pre-test interview or the testing is justified, if, in the examiner's judgment it seems warranted.

Within the R-I tests, of course, there is usually no actual "control" against which responses to the relevant questions can be compared, at least no control similar to that advocated by proponents of the CQ technique. The lack of such a control is believed to make the R-I technique an

"interrogation" capitalizing on the psychological effect of the polygraphic instrument and recordings; R-I tests, then, for reasons to be further explained here are usually considered by proponents of the CQ technique inadequate for making decisions regarding a person's truthfulness or deception based upon the polygraphic recordings exclusively.[20]

Control-Question Technique

Many leading polygraph examiners today distinguish between "interrogation" and polygraphic testing. The major impetus of this change in approach was the "control question" as developed by John E. Reid in 1947.[21] Since Reid's first publication on this topic he and other practitioners have so refined the use of control questions and the procedure used for giving polygraphic tests that it is now believed that polygraphic testing and interrogation must be considered separately. That is, most proponents of the CQ technique believe that polygraphic testing provides a substantially accurate means of determining a person's truthfulness or deception independent of "interrogation"; in fact, interrogation before or during the testing proper is believed detrimental to testing.[22]

---

[20]The discussion concerning the R-I Technique was condensed from: L. Harrelson, Keeler Polygraph Institute Training Guide (Chicago: Keeler Polygraph Institute, 1964).

[21]J. Reid, "A Revised Questioning Technique in Lie Detection Tests," J. Crim. Law and Crim., 37 (1947), 542-547.

[22]J. Reid and F. Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique (Baltimore: Williams and Wilkins, 1966), 177.

th
ex
po
th

by
wh
th
an
in
th
ov
re
be
la
es
ti

th
co
in
no

/

fo
to
11/29/

The C-Q technique consists of two distinct components: the pre-test interview and polygraphic testing. Although some examiners maintain that post-test interrogation is a third component,[23] such a contention seems out of line with the notion that interrogation and polygraphic testing are separate phenonena.

Pre-test Interview.--The pre-test interview as used by proponents of the CQ technique occurs prior to testing, when the examiner discusses with the subject the purpose of the examination, the nature of the polygraphic instrument, and, in general, seeks to prepare the subject for the testing. Unlike the interview used in the R-I technique, however, there is no intensive questioning on the issue at hand. Moreover, during the interview the examiner makes it a point to review with the subject the exact test questions which will be asked, and the subject himself participates in the formulation of these questions. Such participation is considered essential to the functioning of the testing procedure, particularly with respect to the control-questions.

There are, of course, variations among examiners in the way a pre-test interview is conducted. Some examiners conduct a lengthy interview and acquire detailed background information, e.g., medical history, etc., while others do not. Some use specialized interview techniques to become

---

[23] G. Barland and D. Raskin, "The Use of Electrodermal Activity in the Detection of Deception," Prepublication copy to appear in: W. Prokasy and D. Raskin (Eds.), Electrodermal Activity in Psychological Research (New York: Academic Press, in press).

far

in

exa

the

aut

fut

tic

fou

bet

the

the

in

thr

wan

oth

are

pat

you

/

/

/

acti

familiar with behavioral characteristics which may be helpful in making a diagnosis of truthfulness or deception. Some examiners spend a considerable amount of time explaining the nature of the polygraphic instrument, the way in which autonomic responses are used to detect deception, and the futility of trying to beat the test. More detailed information concerning variations in the pre-test interview can be found in Reid and Inbau,[24] Horvath,[25] or Barland and Raskin.[26]

Polygraphic testing.--While there are differences between pre-test interviews in the R-I and CQ procedures, the essential difference between them lies in the nature of the questions asked during polygraphic testing and the manner in which response data are evaluated. During the CQ testing, three basic types of questions are asked: irrelevant, relevant, and control questions, although, as in the R-I technique, other question types may also be used.[27] Irrelevant questions are those used for establishing "normal" or truth-telling patterns; they will deal with such matters as: "Do they call you Joe?" and, "Are you over 21 years of age?" Relevant

---

[24]Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 10-16.

[25]F. Horvath, "Verbal and Nonverbal Clues to Truth and Deception During Polygraph Examinations," J. Pol. Sci. and Adm., 1 (1973), 138-152.

[26]Barland and Raskin, "The Use of Electrodermal Activity in the Detection of Deception," op. cit., 5-8.

[27]Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 18; R. Arther, "The Guilt Complex Question," J. Polygraph Studies, 4 (1969), 1-4.

gue
gat
the
gro
jec
pre
gat
exa
exa
ste
me
see
wil
at
ful
all
int

sub
con
tes
fou
que
que
sec
rep

questions are those which pertain to the matter under investigation, such as "Did you shoot John Doe?" and, "Did you fire the shots that killed John Doe?" Control questions are those growing out of interaction between the examiner and the subject; in general they deal with matters similar to, but of presumed lesser significance than, the offense being investigated. While the interaction between the subject and the examiner determines the exact nature of these questions, an example in burglary-investigation might be: "Did you ever steal anything?" or, "Except for what you have already told me about, did you ever steal anything else?" The examiner seeks to frame these questions in such a way that the subject will answer "no" but will, in all probability, be lying or at least will have some doubt or concern about the truthfulness or accuracy of his answer. After the formulation of all test questions and at the completion of the pre-test interview, polygraphic testing is conducted.

In the polygraphic testing, the examiner asks the subject the previously reviewed irrelevant, relevant and control questions in a series of polygraphic tests. Each test generally consists of about ten or eleven questions, four irrelevant, two control, and four or five relevant questions, and will usually last about three minutes. All questions are asked once during one test, and at about twenty-second intervals. A complete examination consists of the repetition of several of these tests. It is generally agreed

tha

acy

a r

lea

or

tio

test

proc

inte

or l

a ca

is a

expl

stra

actu

"stir

ing t

later

exami

reque

quest

asks

graph

that for an examiner to ascertain with any degree of accuracy the deception or truthfulness of the subject's answer to a relevant test question, that question should be asked at least once on each of two separate tests; sometimes, four or five separate tests may be conducted before a determination of deception is made.[28]

It might be helpful at this point to describe the testing sequence used by many of the proponents of the CQ procedure. Generally, immediately following the pre-test interview, the examiner conducts the first CQ test of 10 or 11 questions, previously reviewed. After this first test, a card (or "numbers") test, or some variation of such a test, is administered. The nature of the card test being fully explained elsewhere,[29] its ostensible purpose is to demonstrate to the subject the efficacy of the "lie-detector"; actually, it is more properly considered one of the many "stimulation" devices or strategies used by examiners employing the CQ procedure. Such strategies will be discussed later.

Following the "card test" the examiner leaves the examination room for a short period, before doing so usually requesting the subject to think carefully about the test-questions while he is out of the room. Upon his return, he asks the subject if there are any questions which concern

---

[28] Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 26-33.

[29] Ibid., 27-28.

him more than others, or if there are any which the subject feels should be re-worded. If not, the examiner then tells the subject that another test will be conducted using the same questions asked in the first test, and in the same order; in other words, the third test is a replicate of the first.

Upon completion of this third test, the examiner briefly reviews the accrued polygraphic recordings and decides if further testing is necessary. It is usually claimed that in some instances, response data contained in the first two control question tests are sufficient to indicate the subject's truthfulness or deception.[30] In the majority of instances, however, further testing is indicated and conducted via one or more of the specialized tests discussed below.

Specialized tests.--1) Mixed Question Test. In most instances of additional testing the first test will be a "mixed question test." In this test the subject is asked the questions of the first two control-question tests but in a different order. The ordering of the questions is flexible, usually based upon the examiner's knowledge of the response-data observed in the prior tests.[31]

2) Silent Answer Test. A specialized test which some examiners have recently incorporated as the fourth test

---

[30] Ibid., 30-37.

[31] Ibid., 30-32.

in

qu

an

el

af

by

(w

pr

he

tes

in

Ord

"ye

sam

The

is

dir

dat

res

the

/

Ans

285

Pra

/

in the series (usually in the position where the mixed question test is placed) has been termed the "silent answer test". Its usefulness has been adequately described elsewhere.[32]

3) The "yes" or Affirmation Test. The "yes" or affirmation test is one in which the subject is instructed by the examiner to answer "yes" to all of the test questions (which, of course, are the same questions already asked on previous tests), including the relevant questions to which he had answered "no" before. The purpose of the "yes" test is to ascertain whether or not the subject is engaging in deliberate attempts to distort his polygraphic recordings. Ordinarly the tracings (response data) obtained during the "yes" test are not interpreted in the same manner or for the same purpose as they are in the tests mentioned previously. The purpose and method of interpretation of the "yes test" is thoroughly discussed in Reid and Inbau.[33]

Stimulation procedures.--Proponents of the CQ procedure have developed various strategies to clarify response data; that is, these strategies are used not only to augment responsiveness to testing but, more importantly, to direct the subject's attention (or psychological set) to those test

---

[32]F. Horvath and J. Reid, "The Polygraph Silent Answer Test," J. Crim. Law, Crim., and Pol. Sci., 63 (1972), 285-293.

[33]Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 32.

ques
bein
stra
lyin
form
answ
exam
take
less
inst
by e
subj
effe
reco
to r
then
thre
CQ T
resp
the
the
dis
an i
test
the

questions which constitute the greatest threat to his well-being; presumably, for persons telling the truth these strategies augment responses to control questions; for those lying, to relevant questions. Such strategies may take the form of specialized tests, e.g., the "card test", "silent answer test", etc., or, may consist of various forms of examiner-subject interaction. Regardless of which form they take, however, these strategies are considered to be much less direct than ordinary interrogational devices. For instance, when compared to direct questioning, implications by either verbal or nonverbal communication, concerning the subject's polygraphic records are considered to be much more effective and less apt to adversely affect polygraphic recordings, i.e., cause a person to respond beyond the normal to relevant test questions when he is telling the truth to them. Perhaps an example would clarify this point.

Assume that an examiner has conducted a series of three tests with a subject (CQ-Test One, a card test, and CQ Test Three -- a repetition of Test One) and feels that the responses are too ambiguous to permit accurate appraisal of the subject's truthfulness in answer to the relevant questions -- the responses to the control questions cannot be clearly differentiated from those to the relevant questions. In such an instance, the examiner may feel that a mixed-question test is warranted. Before conducting such a test he may ask the subject if any particular test questions concern him more

than the others; while doing so he implies that the testing is not "clear" at this point. Further he may tell the subject that he would like to conduct another test but that before he does so, he wants to be certain that the subject clearly understands all of the test questions so far asked and is certain that he has answered all of them truthfully. The examiner may then carefully re-read all of the test questions, requesting answers as he does so. He then asks the subject something like: "Are you certain that you understand all of these questions?" "Is there any answer you have given that may not be the complete truth?" When the subject acknowledges he has answered the questions truthfully and that he understands all of them, the examiner explains how the next test is to be conducted, i.e., the same questions will be asked in a different order than they were asked on prior tests, and then proceeds with the testing.

The various strategies used by examiners to "stimulate" subjects are too numerous to detail here. It should be noted, however, that the strategies are rather indirect in nature; they are not accusatory and do not usually make reference to particular test-questions, and most importantly, they presumably make a significant contribution to the functioning of the CQ procedure.[34]

---

[34]J. Reid, "Stimulation Technique Outline," undated, unpublished manuscript supplied by J.E. Reid and Associates, Chicago; C. Klump, "Principles of Controlled Stimulation" (paper presented at American Academy of Polygraph Examiners, Eighth Annual Seminar, Washington, D.C., Sept., 1961).

While the general testing procedure outlined above is representative of that used by many field examiners employing the CQ procedure, there are other specialized tests and other variations of the procedures. Some of these variations concern the number of individual tests which will be conducted during an examination, the organization of the tests, the order of questions within tests, and the procedure followed by the examiner during the break between tests. For a more thorough discussion of these variations see Reid and Inbau,[35] Barland and Raskin,[36] or Backster.[37]

Regardless of the various administrations of the CQ test, its proponents argue that control questions imbedded within the series provide a better tool for assessment of a person's truthfulness or deception to relevant issues than does the R-I procedure. The variations do not imply unstructured procedure, however, each variation being controlled by its particular rules for conducting examinations. Presumably, once informed of each others' rules, examiners using the different procedures of examination can evaluate each other's results.

Peak of Tension Testing

A type of testing infrequently encountered in field settings is the POT (peak of tension) test. Although the

---

[35]Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 10-36.

[36]Barland and Raskin, "The Use of Electrodermal Activity in the Detection of Deception," op. cit., 13-17.

[37]C. Backster, Standardized Polygraph Notepack and Technique Guide (New York: Backster Research Foundation, 1969).

pri
of
ing
sta

the
POT
all
For
may
in a
wea
inc
to
to
bei
exa

Sim
Sev
Par
any

principle behind this test is often relied on by proponents of both the R-I and CQ procedures, especially in the ordering of questions in the test series, the POT is not a standard part of either of these procedures.

Arthur has termed the two general forms of the POT tests, the "searching" test and the "known-solution" test.[38] The searching POT consists in the asking of a series of similar questions, usually, with specific focus, such as to locate a murder weapon, etc. For example, a subject tested by control-question type testing may give the examiner reason to think that he is in fact implicated in a certain murder and further has hidden or discarded the murder weapon. Under these circumstances the searching POT test would include a series of questions such as: "Do you know if the gun used to kill John Jones is under water?", "Do you know if the gun used to kill John Jones is buried in the ground?", etc., such questions being asked throughout a number of individual tests until the examiner feels he has determined the location of the murder weapon.[39]

On the other hand, the known-solution POT test, while similar to the searching test consisting of a series of about seven questions presupposes that the examiner is aware of particular details of a crime of which the subject denies any knowledge. For example, the examiner may know that in a

---

[38] R. Arther, "Peak of Tension: Basic Information," J. Polygraph Studies, 1 (Jan.-Feb., 1967), 4.

[39] See: Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 37-40; R. Arther, "Peak of Tension: Examination Procedures," J. Polygraph Studies, 5 (July-Aug., 1970), 1-4.

cer

sto

suc

"Do

the

usu

pre

sta

far

gra

wil

rea

Pur

as

the

not

in

lie-

Ste
cer

gra

Dete
logi
a. S
Holt

certain burglary two hundred dollars in quarters has been stolen. The subject is then asked a series of questions such as: "Do you know if dimes were stolen in X burglary?", "Do you know if nickels were stolen in X burglary?", etc., the critical question, in this case the one about the quarters, usually placed in the fourth position in the series.

Regardless of the type of POT test employed, interpretation of the polygraphic records thus obtained is standard. It is assumed that if a subject is in fact familiar with the critical item in the series, the polygraphic recordings (especially the "cardio" and GSR tracings) will appear to "peak" at the critical item or will show a reaction of the greatest magnitude at the "critical" item. Further ramifications of the POT test and its interpretation, as well as necessary precautions in its use are recorded in the literature.[40] For the purposes of this study it should be noted that in the POT test examiners rely heavily on reactions in electrodermal activity as indications of deception.[41]

Contrary to some writings,[42] the POT test is not a lie-detection "technique" in the sense that the control

---

[40] R. Arther, "Peak of Tension: Dangers," J. Polygraph Studies, 2 (March-April, 1968), 1-4; Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op.cit., 37-40.

[41] Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 219-225.

[42] M. Orne, R. Thackray and D. Paskewitz, "On the Detection of Deception: A Model of the Study of the Physiological Effects of Psychological Stimili," N. Greenfield and R. Sternbach (Eds.), Handbook of Psychophysiology (New York: Holt, Rinehart and Winston, 1972), 743-780.

ques

Rath

test

ques

used

spec

is l

are

are

part

R-I

exam

"mix

and

with

Eval
Reco

if e

ing

vasc

cp.

at A
Hous

question and relevant-irrelevant procedures are techniques. Rather, the POT is merely a specialized type of polygraphic test normally used only after testing by either the control-question or relevant-irrelevant procedures; the POT test is used to determine if a given person has "guilty knowledge" of specific details of a particular offense.[43] Hence, its use is limited to those types of offenses where such details are evident. On the other hand, the CQ and R-I procedures are diagnostic techniques not predicated on awareness of particular details of an offense. Generally, these CQ or R-I techniques can be administered in a variety of ways, the examiner having at his disposal the specialized "card test", "mixed question test", "yes test", "silent answer test",[44] and "yes-no test",[45] and others, all of which can be used within the framework of either the CQ or R-I technique.

Evaluation of Polygraphic Records

Visual inspection technique.--Field examiners rarely, if ever, employ strictly objective measurements in interpreting the significance of response-data, changes in cardiovascular, respiratory, or GSR tracings recorded polygraphically.

---

[43] R. Arther, "Peak of Tension: Basic Information," op. cit., 4.

[44] See page 28.

[45] R. Golden, "The Yes-No Technique" (paper presented at American Polygraph Association Seminar, August, 1969, Houston, Texas).

Rat
gen
ana
usu
of
the
abl
in
as
may
do
res
the
fie
cha
dec
car
be
/
gra
/
Res
and
/ec
C.
ter
J.
Res
aut

Rather, visual inspection techniques, progressing from a general appraisal of all records (tests) down to particular analysis of reactions to particular test questions, are usually performed. Generally, changes - extent and duration of cardiovascular, respiratory, or GSR response - in any of the recorded parameters are evaluated according to specifiable criteria for each parameter as set forth in texts,[46] or in training manuals.[47] Such criteria, however, serve only as guidelines, since the "deception-responses" of one person may not be those of another. In other words, field examiners do not claim that any particular response, or pattern of responses is pathognomic of lying, only that changes from the "normal" for any given person may indicate deception.[48]

Some writers have over-generalized the evaluation of field-derived polygraphic records to the point where any change from pre-stimulus levels is said to be indicative of deception. While it is true that polygraphic records indicate any changes from pre-stimulus levels, such changes must be considered both quantitatively and qualitatively, they

---

[46]Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 41-50.

[47]C. Backster, Tri-Zone Polygraph (New York: Backster Research Foundation, 1969).

[48]See: C.N. Joseph, "Analysis of Compensatory Responses and Irregularities in Polygraph Chart Interpretation," Academy Lectures on Lie Detection, V. Leonard (Ed.) (Springfield, Ill.: C. C Thomas, 1957), 93-99; P. Trovillo, "Deception Test Criteria," J. Crim. Law, Crim. and Pol. Sci., 33 (1942), 338-358; J. Reid, "Interpretation of Truth and Deception in Polygraph Test Records," undated, unpublished manuscript supplied by author.

cannot be summarily assumed indications of deception. Consider record-evaluation in the control-question technique, for example. Simply stated, responses in the polygraphic parameters which occur more consistently over a series of tests and which are of a greater intensity to control-questions than to relevant questions, indicate truthfulness to the relevant questions. Conversely, responses of a consistently greater intensity to the relevant question than to the control questions suggest deceptiveness regarding the relevant questions. The key points in this vastly over-simplified description, are that any changes have little significance unless they occur consistently, and even then they are not significant until compared with other changes.

Numerical evaluation technique.--One of the noteworthy variations in evaluation of polygraphic recordings is a numerical scoring system developed by Backster, a well-known field examiner.[49] In this system examiners assign a number ranging from -3 to +3 to reflect the perceived difference between responses to control and relevant question pairings for each of the physiological parameters recorded; the magnitude and direction of the numbers assigned to such comparisons forms the basis for decision-making. For example, the examiner pairs relevant and control

---

[49]Backster, Tri-Zone Polygraph, op. cit., 14.

questi

questi

the re

from

is ass

respor

there

assigr

each

parame

are th

tive

points

Total

incon

cal s

such

in eva

only

someti

genera

no pr

subjec

Polyc

exampl

questions and then observes whether or not a particular question in each pair provokes outstanding response. If the response is greater to the relevant question, a number from -1 to -3, depending upon the extent of the difference, is assigned. On the other hand, if the control-question response is greater, a number from +1 to +3 is assigned; if there is no difference between the paired responses, a 0 is assigned. Such a procedure is carried out separately for each control/relevant-question pair for each physiological parameter of all the tests administered. The numbers assigned are then added; a positive total greater than 5 and a negative total less than 5 usually are established as "cut off" points to indicate truthfulness and deception, respectively. Total scores ranging between +5 and -5 are usually considered inconclusive.

There are some disadvantages apparent in the numerical scoring system: (1) It is possible that scoring data in such a way filters out recorded trends which might be useful in evaluation. (2) It assumes that response-data are the only indices of deception. In actuality, deception is sometimes indicated not so much by specific response as by generally abnormal or erratic recordings. (3) It makes no provision for artifacts deliberately produced by some subjects.[50] Within its limits, however, the numerical-scoring

---

[50]See: Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., for specific examples of these three phenomena, 53-124, 185-218.

sys

use

Dis
of

maj

dif

gra

obt

bef

of

dia

are

the

two

con

ther

ext

the

tiv

val

dec

Eva
tio

system appears to be highly reliable and an especially useful research tool.[51]

Discussion and Summary of Field Procedures

It should be evident from this discussion of the major procedures used in the field, that it is extremely difficult to separate the polygraphic testing or the polygraphic records themselves from the procedure used in obtaining them. That is, the examiner-subject interaction before and during polygraphic testing is an integral part of the procedure; one must view field "lie-detection" as a diagnostic technique whether or not R-I or CQ procedures are considered. The most prominent distinction between these procedures seems to be that if one were to place these two "lie-detection" procedures on a subjective-objective continuum, proponents of the CQ procedure would place themselves more to the right, or towards the objective extreme, of the continuum. It is clear that they believe the use of control questions a necessary basis for objectivity, that the polygraphic recordings themselves are highly valid and reliable indicators of a person's truthfulness or deception.

---

[51] G. Barland, "The Reliability of Polygraph Chart Evaluations" (paper presented at American Polygraph Association Seminar, Aug. 15, 1972, Chicago, Ill.).

Laboratory Lie-Detection: Procedures

Laboratory studies of lie-detection usually involve either a guilty-person or a guilty-information paradigm, the two not mutually exclusive.[52] Following the guilty-person paradigm, a mock crime is contrived; the task of the examiner is to employ lie-detection apparatus to determine which of a given group of subjects committed the crime, which were accomplices, and which were free of any complicity. This testing is closely akin to the relevant-irrelevant tests used in field settings; control-question testing, somewhat similar to that used by field examiners, is recorded in only one laboratory study.[53] In the guilty-information paradigm the subject is instructed to lie about a card, number, or some other item he selects from a group of such items; the examiner's task is to determine which item was selected, hence, the process can be generally viewed as a "peak-of-tension" test.

One of the noteworthy variations of the two laboratory paradigms is termed the "guilty-knowledge technique", originally reported by Lykken.[54] Using this technique, subjects assigned to groups who may have committed one or more, or no mock crimes are interspersed among irrelevant, or non-critical

---

[52] Orne, et al., "On the Detection of Deception: A Model for the Study of the Physiological Effects of Psychological Stimuli," op. cit., 775.

[53] G. Barland, "An Experimental Study of Field Techniques in Lie Detection" (unpublished M.A. Thesis, University of Utah, 1972).

[54] D. Lykken, "The GSR in The Detection of Guilt," J. Appl. Psych." 43 (1959), 385-388.

items

aware

physi

infor

for q

to th

hence

"inno

varia

exami

that

to de

cant

as ev

as su

ity a

tory

respo

of co

and a

gener

is c

Minn
Repo

items. It is presumed that those guilty of the crimes, aware of certain information about them, will give augmented physiological responses to test items pertaining to such information. And, therefore, in a series of such tests (or questioning) guilty persons could be expected to respond to the critical items more often than would innocent persons; hence, some estimate of whether a person is "guilty" or "innocent" is possible.

The guilty-knowledge technique appears to be a variation of the known-solution POT test used by field examiners. Lykken, however, argues otherwise, believing that it is a "very different thing to use the polygraph to determine whether the subject can identify the significant alternative, than to use autonomic arousal or "tension" as evidence that the subject is lying."[55]

Typically, laboratory studies use college students as subjects, employ only a measure of electrodermal activity as the physiological (dependent) variable, use laboratory personnel as examiners, and, most often analyze response data by some objective technique. These factors, of course, tend to insure rigorous statistical analysis and adequate control over data-collection although the generalization of results is greatly restricted. Moreover, it is clear that laboratory research approaches lie-detection

---

[55]D. Lykken, Psychology and The Lie Detector Industry (Minneapolis: Department of Psychiatry, Univ. of Minnesota, Report No. PR-74-1, January 25, 1974), 14.

in a

sub

Fie

the

bet

sou

res

soc

exc

her

res

pra

tod

def

sig

in a manner quite different from that in the field; examiner-subject interaction seldom has a very dramatic impact.

The Validity of Lie-Detection Field Procedures

The validity of field lie-detection procedures, i.e., the accuracy with which lie-detection can discriminate between truthful and lying persons, has been a constant source of debate between field practitioners, laboratory researchers and others concerned with this problem and its social implications.[56] Because there are already available excellent discussions of this topic,[57] the presentation here will be relatively brief, only the most prominent research results and related problems discussed.

As noted previously, many of the early lie-detection practitioners used procedures and instrumentation which by today's standards appear unsophisticated. In spite of this deficiency, however, there are numerous reports of impressive validity. Bennussi, for instance, claimed that he was

---

[56] See, for example: U.S. Congress, House, Subcommittee of the Committee on Government Operations, Use of Polygraphs as "Lie Detectors" by the Federal Government, Hearings, 88th Congress, 2nd Sess., and 89th Congress, 1st Sess., Parts 1-6 (Washington, D.C.: U.S. Government Printing Office, 1964-1966).

[57] See: S. Abrams, "Polygraph Validity and Reliability: A Review," J. Forensic Sciences, 18 (1973), 313-326; Barland and Raskin, "The Use of Electrodermal Activity in the Detection of Deception," op. cit., 1-62; J. Orlansky, An Assessment of Lie Detection Capability (Declassified Version), Tech. Rep. 62-16 (Arlington, VA: Inst. for Defense Analyses, Res. and Eng. Support Div., July 1964), 6-17; Orne, et al., "On the Detection of Deception: A Model for the Study of the Physiological Effects on Psychological Stimuli," op. cit., 743-780.

able to successfully detect liars by evaluating the respiration-inspiration-expiration ratio; the ratio was greater before truth-telling than after, and greater after lying than before.[58] Marston claimed greater success with discontinuous systolic-blood pressure as a test of deception, and reportedly could discriminate between truth-tellers and liars with an accuracy of 96 percent.[59] In contrast, Summers rejected the value of both respiration and blood pressure and relied on a measure of electrodermal activity. He claimed 98 percent success in discriminating between truth-tellers and liars in the laboratory and 100 percent success when dealing with actual criminal suspects.[60]

Benussi, Marston, and Summers, of course, did not use a polygraph -- but a single-channel recorder. Larson and Keeler, using polygraphic recording equipment, claimed to have accuracy rates varying between 90 and 100 percent.[61] Inbau and Reid claimed an accuracy of 95.6 percent in their initial report on this topic.[62] Likewise, Arther, estimating

---

[58]Benussi, "On the Effects of Lying on Changes in Respiration," cited by Trovillo, "A History of Lie Detection," op. cit., 870.

[59]Marston, "Systolic Blood Pressure Symptoms of Deception," op. cit., 123

[60]Cited by Trovillo, "A History of Lie Detection," op. cit., 108.

[61]Larson, Lying and Its Detection, op. cit., 405-416; Keeler, "A Method For Detecting Deception," op. cit., 38-52.

[62]F. Inbau and J. Reid, Lie Detection and Criminal Interrogation (Baltimore: Williams and Wilkins, 1953), 110-113.

from the results of a five-year study, reported an accuracy of over 96 percent with a 3 percent margin of inconclusive determinations and a 1 percent margin of maximum error; he reported that his known error was actually less than .0005.[63]

In view of such favorable reports of the accuracy of lie-detection in the field setting, it is logical to question how well such reports stand up in objective assessment. Inbau and Reid's early claim of 95.6 percent accuracy had been arrived at by adding instances in which examiners made judgments of lying (31.1 percent) or truth-telling (64.5 percent) in a number of cases. The remaining 4.4 percent of the judgments were inconclusive and the reported error was 0.0007 percent, which was later pointed out as being in arithmetical error to be corrected to 0.07 percent.[64]

The verification of the Inbau and Reid data rested on confessions made by the persons tested. However, only 486 out of 1334 (36.4 percent) persons who were judged to be liars actually confessed, and only 11.7 percent of the judgments made on the truth-tellers could be verified. Thus, Inbau and Reid defined accuracy as the percentage of cases in which the examiner made a determination of either lying or truth-telling irrespective of actual verification. This

---

[63]R. Arther and R. Caputo, Interrogation For Investigators (New York: W.C. Copp, 1959), 214.

[64]Orlansky, An Assessment of Lie Detection Capability, op. cit., 13.

wa

be

in

su

ap

ma

st

in

to

th

es

La

ga

in

re

Ha

Wh

---

bi
Co
(1

Po
07
Ma
26

3.

was an unusual interpretation of "accuracy" and has since been strongly criticized.[65] Many other field examiners have interpreted their accuracy in the same manner and are thus subject to the same criticism.

Field practitioners have also reported studies approaching the question of validity in a more acceptable manner. It is unfortunate that the majority of these studied are quite old and either did not employ polygraphic instrumentation[66] or did not use procedures commonly used today.[67] Moreover, many field reports of the accuracy of the polygraph rely on anecdotal evidence which, while interesting, is not an acceptable method of determining validity. Larson, for instance, reported an investigation in which he gave polygraphic tests to a number of girls living together in a large hall in order to determine which of them was responsible for a series of thefts amounting to about $600.00. He reportedly was able to "clear" all but one of the girls who subsequently confessed; thus, an accuracy of 100 percent

---

[65] Orlansky, An Assessment of Lie Detection Capability, op. cit., 11; R. Sternbach, L. Gustafson, and R. Colier, "Don't Trust the Lie Detector," Harv. Bus. Rev., 40 (1962), 130.

[66] W. Summers, "Science can get the Confession," Fordham Law Rev., 8 (1939), 334-354; R. MacNitt, "In Defense of the Electrodermal Response and Cardiac Amplitude as Measures of Deception," J. Crim. Law and Crim., 33 (1942), 266-275.

[67] V. Lyon, "Deception Tests with Juvenile Delinquents," J. Gen. Psych., 48 (1936), 494-497.

was

is

per

50

inf

det

gir

suc

tes

is

cri

and

exc

par

in

in

th

co

in

i

i

a

c

was claimed. The problem with such an "accuracy", of course, is that the group of girls tested contained only one guilty person, the likelihood of being innocent or guilty was not 50 percent. Moreover, as Larson points out, the factual information available was sufficient to enable him to determine in advance of the testing that certain of the girls were more likely to have been "guilty" than others; such information could easily have influenced the polygraph testing.[68]

The most enlightening validity-study reported to date is by Bersh; he drew a random sample of cases from a pool of criminal investigations carried out by the military services and submitted complete dossiers of all evidence in the cases, except for any reference to polygraphic examinations, to a panel of four military lawyers. All evidence was reviewed independently by the lawyers and determinations of guilt or innocence were made irrespective of legal technicalities; these determinations were then used as the criteria for comparison with the examiners' judgments. In those instances in which all four lawyers agreed on a subject's guilt or innocence, the judgments of the polygraphic examiners were in agreement with the lawyers 92.4 percent of the time. When a majority determination by the lawyers was used as the criterion of guilt or innocence, agreement with the polygraphic

---

[68] Larson, "Modification of the Marston Deception Test," op. cit., 395-396.

exam

and

agre

and

not

is t

infl

as b

out:

the

poly

Acco

the

grap

sely

the

tion

grap

crim

to h

the

corr

Judg

examiners' judgments was 74.6 percent; and, when unanimous and majority decisions were combined, an 87.5 percent agreement obtained.[69]

While the Bersh study is of considerable interest and may represent a very useful approach to validity, it is not without some serious deficiencies. Foremost among these is the fact that the examiners' judgments may have been influenced as much by the polygraphic recordings themselves as by their knowledge of other information; as Bersh points out: "No attempt was made to disentangle the influence of the polygraph examination and record from that of the extra-polygraph sources of information available to the examiner."[70] Accordingly, Bersh's results bear only upon the validity of the examiners' judgment, not upon the validity of the polygraphic procedure or of the polygraphic recordings themselves.

In an attempt to disentangle the judgments made on the polygraphic recordings from those made on other information, Holmes submitted to a group of six experienced polygraphic examiners the recordings of 32 persons involved in criminal investigations. Twenty of the persons were known to have lied during their examination, twelve to have told the truth. The criteria used for such verification were corroborated confessions.

---

[69] P. Bersh, "A Validation Study of Polygraph Examiner Judgments," J. Appl. Psych., 53 (1969), 399-403.

[70] Ibid., 400.

pol
tru
wer
exa
mat
ter
opi
acc
Hol
lia
ver
con
gra
the
the
sul
—
Int
Sec
gra
Law
and
Exa
37
of
Exa
De
rep

The examiners were initially asked to evaluate the polygraphic recordings and to identify which were those of truth-tellers and which of liars. Correct determinations were made, on the average, 75 percent of the time by the examiners. When Holmes gave the examiners additional information about the subjects, such as their behavioral characteristics during the testing, investigators' reports and opinions, and witnesses' accounts of the offenses, etc., accuracy rates increased to 83 percent overall. Moreover, Holmes found that errors more often favored the lying persons, liars, more often judged to be truth-tellers than vice versa.[71] Unfortunately, Holmes did not report details concerning the testing procedure used in obtaining the polygraphic records or the experience levels and the nature of the training of the examiners who evaluated the records; these variables could have significantly affected the results.[72]

---

[71] W. Holmes, "The Degree of Objectivity in Chart Interpretation," Academy Lectures on Lie Detection, II, V. Leonard (Ed.) (Springfield, Ill.: C.C Thomas, 1958), 62-70.

[72] F. Horvath and J. Reid, "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," J. Crim. Law and Crim., and Pol. Sci., 62 (1972), 276-281; F. Hunter and P. Ash, "The Accuracy and Consistency of Polygraph Examiner's Diagnoses," J. Pol. Sci. and Adm., 1 (1973), 370-375; A. Suzuki, "An Analysis of Relative Effectiness (sic) of the Physical Indices and the Influence of Polygraph Examiner's Experience Upon Judgment of Polygraph Records in Detection of Deception," Japanese Journal, Title unknown, reprint supplied by author, 21 (1968), 51-59.

Reported in the literature are several other studies utilizing a design somewhat similar to that used by Holmes. However, for reasons which will be discussed at a later point, these studies are more appropriately viewed as reliability rather than validity studies.

While the validity of field lie-detection procedures is a crucial concern, it is clear that as yet the evidence supporting extremely high accuracy in the field is inconclusive. The major reason for the lack of supporting evidence, of course, is that there is no completely adequate ground-truth criterion with which examiners' judgments can be compared. The criteria which have been or can be used, such as confessions, independent evaluations of extrapolygraphic information, and the outcome of judicial proceedings, do not establish with certainty a person's actual truthfulness or deception.[73] And, since procedures used in giving polygraphic examinations are, in essence, diagnostic procedures, it is difficult to separate the influence of the examiner's interaction with the subject from the polygraphic recordings themselves; that is, the recordings are not necessarily independent of the examiner's attitudes, behavior, and information concerning the subject's involvement

---

[73]For a discussion of the problems associated with the use of confessions as a ground-truth criterion see: H. Dearman and B. Smith, "Unconcious Motivation and the Polygraph Test," Amer. J. Psych., 119 (1963), 1017-1021; R. Ferguson, The Scientific Informer (Springfield, Ill.: C. C Thomas, 1971).

in t

has

one

the

othe

Such

been

Labo

derm

will

is w

derm

phen

dece

rece

inve

Ruch

the

Psy

Phe

Pra

P.

the

Cla

in the offense under investigation. For this reason it has been argued that the proper approach to validity is one which compares the validity of the various aspects of the polygraphic technique separately and collectively against other methods of determining truthfulness or deception.[74] Such an approach is quite reasonable but as yet has not been reported in the literature.

Laboratory Procedures

Because laboratory research typically uses electrodermal activity to indicate deception, the discussion here will be restricted to the validity of this phenomenon. It is well established that during the early 1900's electrodermal activity was known to be associated with "psychic phenomena" such as lying.[75] However, attempts at detecting deception with electrodermal activity probably did not receive full impetus until the 1930's. At that time many investigators reported substantial success with the method. Ruckmick, using the guilty-information paradigm reported that

---

[74] M. Orne, "Implications of Laboratory Research for the Detection of Deception," Polygraph, 2 (1973), 169-199.

[75] See: C. Landis, "Electrical Phenomenon of the Skin," Psych. Bull., 29 (1932), 693-752; C. Landis and H. DeWick, "The Electrical Phenomenon of the Skin (Psychogalvanic Reflex), Psych. Bull., 26 (1929), 64-119; J. Larson, "The Cardio-Pneumo Psychogram and Its Use in the Study of Emotions, with Practical Applications," J. Exp. Psych., 5 (1922), 323-328; F. Peterson and C. Jung, "Psycho-Physical Investigations with the Galvanometer and Pneumograph in Normal and Insane Individuals," Brain, 30 (1907), 153-218.

a 6

and

wor

four

elir

let

para

ada

coul

ada

trar

no i

resp

util

the

With

corr

Mock

J. A

as a
346-

cit.

a 66 percent detection rate was achieved with numbered cards, and using the same paradigm with a series of three-letter words, achieved 78 percent correct judgments. Moreover, he found that if the scores of an inexperienced evaluator were eliminated, an 83 percent accuracy was achieved for the three-letter words.[76] Geldreich, also using the guilty-information paradigm with decks of cards, claimed that by "fatigue-adapting" a group of subjects to non-critical cards he could improve detection rates from 74 percent for a non-adapted group to 100 percent for an adapted group.[77]

Fatigue-adapting, Geldreich concluded, shunted extraneous stimuli to non-critical items, although there is no indication that he also controlled for differential response-capabilities between groups prior to his experiment.

Summers, in what is perhaps the earliest attempt to utilize the guilty-person paradigm, claimed to have improved the galvanometer and the technique used for scoring responses. With his Fordham Pathometer he reported that he was able to correctly detect "guilty", "innocent" and "accomplices" in mock crimes 98 percent of the time.[78] He apparently

---

[76]C. Ruckmick, "The Truth About the Lie Detector," J. App. Psych., 22 (1938), 50-58.

[77]E. Geldreich, "Studies of the Galvanic Skin Response as a Deception Indicator," Trans. Kans. Acad. Sci., 44 (1941), 346-351.

[78]Summers, "Science can get the Confession," op. cit., 334-354.

att

"la

on

cas

tha

in

onl

per

tic

in

ha

acc

th

El

ga

se

co

gu

Th

ne

to

th

/

Of

sp

ci/

attributed his failure to achieve 100 percent accuracy to "laboratory conditions."[79] However, MacNitt, commenting on the accuracy of electrodermal response in experimental cases, "mock crimes", and actual field conditions, reported that his interpretations were 99 percent accurate whereas, in guilty-information situations he was able to achieve only a 75 percent accuracy.[80] Hence, Summers' failure at perfection may not have been due to only laboratory conditions.

While the early reports of nearly perfect accuracy in detecting deception with electrodermal actitivity measures have not, in general, been confirmed in more scientifically acceptable experiments, recent investigations have shown that detection rates far beyond chance can be achieved. Ellson, Davis, Saltzman and Burke for instance, using the galvanic skin response (GSR) as an indicator, conducted a series of lie-detection experiments. Initially, they were concerned with the accuracy of GSR responses in detecting guilty-information and the effect of repetition on accuracy. Their results indicated an 80 percent accuracy-rating for mere detection of information; this figure dropped slightly to 70 percent in one repetition of the experiment. When they repeated their experiment to test for the effect of the

---

[79]Cited by: Trovillo, "A History of Lie Detection," op. cit., 108.

[80]MacNitt, "In Defense of the Electrodermal Response and Cardiac Amplitude as Measures of Deception," op. cit., 266-275.

sub

com

res

was

Oth

Ell

per

lis

GSF

by

and

the

cor

per

Re

in
19

in
Ef
De
J.
no
Cl

Cl

Te
25

subject's knowledge of successful lying on a first trial compared to a second trial, they found that by combining the results of their two experiments an accuracy of 79 percent was achieved against a chance-expectancy of 17 percent.[81] Other studies have substantially confirmed the findings of Ellson et al., in both the guilty-information[82] and guilty-person paradigm.[83]

Using the guilty-knowledge technique and by establishing an arbitrary cutoff point for objective analysis of GSR reactions, Lykken was able to correctly classify subjects by group 89.9 percent of the time and to identify the guilty and the innocent 93.9 percent of the time.[84]

In a follow-up study to assess the effects of faking the guilty-knowledge technique, Lykken achieved 100 percent correct classification of subjects who concealed items of personal information.[85] Studies by other investigators have

---

[81]Ellson, David, Saltzman and Burke, A Report of Research on Detection of Deception, op. cit., 11.

[82]D. Van Buskirk and F. Marcuse, "The Nature of Errors in Experimental Lie Detection," J. Exp. Psych., 47 (1954), 187-190.

[83]Barland, "An Experimental Study of Field Techniques in Lie Detection," op. cit.; L. Gustafson and M. Orne, "The Effects of Task and Method of Stimulus Presentation on the Detection of Deception," J. App. Psych., 48 (1964), 383-387; J. Kubis, "Experimental and Statistical Factors in the Diagnosis of Conciously Suppressed Affective Experiences," J. Clin. Psych., 6 (1950), 12-16.

[84]Lykken, "The GSR in the Detection of Guilt," op. cit., 385-388.

[85]D. Lykken, "The Validity of the Guilty Knowledge Technique: The Effects of Faking," J. App. Psych., 44 (1960), 258-262.

al

gu

wh

The

exa

and

tha

is

rea

lab

Dec

der

fie

pur

far

car

"Gu
Mea
"Va
Mot

Pol

also reported varying degrees of success using GSR in the guilty-knowledge technique.[86]

Comparison Of The Validity Of Field To Laboratory Lie Detection

There is general agreement that lie-detection, whether in the field or laboratory, is a valid procedure. The question, is whether or not it is as valid as field-examiners claim. As yet, the evidence is not conclusive, and it may never be. But field-practitioners often claim that given the conditions of their situation, lie-detection is more valid than it is in the laboratory. Several major reasons have been offered for the dissimilarity between laboratory findings and claims of field-examiners.

Deception Indices

In spite of the typically high accuracy of electro-dermal measures in the laboratory, examiners who work in field settings almost universally agree that for their purposes cardiovascular and respiratory measurements are far more effective.[87]

Early accounts of the accuracy of lie-detection using cardiovascular activity reported fairly high accuracy-rates

---

[86] G. Ben Shakhar, I. Lieblich and S. Kugelmass, "Guilty-Knowledge Technique: Application of Signal Detection Measures," J. App. Psych., 54 (1970), 409-413; P. Davidson, "Validity of the Guilty Knowledge Technique: The Effects of Motivation," J. App. Psych., 52 (1968), 62-65.

[87] Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 40.

eve

cor

tel

Mar

lia

rep

see

vas

---

in
R.
5 (
Pre
6 (
of

Dec

Dec
(19

bil
Art
pre
(60
Rug
Dir
Ire
Bir
Uni
Lie
Men
Sur
Law
I.
ter
(19
Phy
Phy
Ass
Par
412
Nov

even in mock crimes.[88] Chappell and Matthew, claimed a correct discrimination rate of 87 percent between subjects telling the truth and lying about details of a mock crime.[89] Marston reported a 94 percent correct classification of liars and truth-tellers.[90] Recent investigators have not reported results as outstanding as these; in fact, recent evidence seems to indicate that for laboratory purposes at least, cardiovascular activity is inferior to electrodermal measures.[91]

---

[88] N. Chappell and N. Matthew, "Blood Pressure Changes in Deception," Arch. Psych., 17 (1929), 1-39; C. Landis and R. Gullette, "Studies of Emotional Reactions," J. Comp. Psych., 5 (1925), 221-253; C. Landis and L. Wiley, "Changes of Blood Pressure and Respiration During Deception," J. Comp. Psych., 6 (1926), 1-19; W. Marston, "Systolic Blood Pressure Symptoms of Deception," op. cit., 117-163.

[89] Chappell and Matthew, "Blood Pressure Changes in Deception," op. cit.

[90] W. Marston, "Psychological Possibilities in the Deception Test," J. Amer. Inst. of Crim. Law and Crim., 11 (1921), 551-570.

[91] J. Kubis, Studies in Lie Detection: Computer Feasibility Considerations, Tech. Report 62-205 (Arlington, Va.: Armed Services Technical Information Agency, June, 1962), prepared for Air Force Systems Command, contract No. AF 30 (602)-2270, Project No. 5534, Fordham University, 1962; S. Kugelmass, Effects of Three Levels of Realistic Stress on Differential Psychological Reactivities, Tech. Report 63-61 (report prepared for Air Force Office of Scientific Research, European Office, Aerospace Research, U.S. Air Force, Hebrew University of Jerusalem, Isreal, Aug. 1963); S. Kugelmass, I. Lieblich, A. Ben-Ishai, A. Opatowski and M. Kaplan, "Experimental Evaluation of Galvanic Skin Response and Blood Pressure Change Indices During Criminal Interrogation," J. Crim. Law, Crim., and Pol. Sci., 59 (1968), 632-635; S. Kugelmass I. Lieblich, "Effects of Realistic Stress and Procedural Interference in Experimental Lie Detection," J. App. Psych., 50 (1966), 211-216; R. Thackray and M. Orne, "A Comparison of Physiological Indices in Detection of Deception," Psychophysiology, 4 (1968), 329-339; R. Violante and S. Ross, Research on Interrogation Procedures (Interim Report, prepared for U.S. Navy, Office of Naval Research, Contract Nonr. 4129(00), Stanford Research Institute, Menlo Park, California, Nov. 1964).

ag

vi

di

ca

po

re

in

wh

Le

to

in

su

In

in

/

Pe

Ra

Te

C.

ff

In spite of the fact that early investigators disagreed on the relative values of either cardiovascular activity or respiration as indicators of deception most of them did find that respiratory measures were fairly good indicators of deception.[92] This is a particularly interesting point since almost all recent investigations have found respiratory measurement to have little, if any, significance in the detection of deception in the laboratory, at least when compared to other physiological parameters.[93]

Level of Subject Affect

One of the reasons that cardiovascular and respiratory activity may be less effective in indicating deception in the laboratory than is electrodermal activity, is that in such settings the level of affect is lower than in real-life. In order to investigate this possibility many laboratory investigators have employed stress and motivational devices

---

[92] Benussi, "On the Effects of Lying on Changes in Respiration," op. cit.; Burtt, "The Inspiration-Expiration Ratio During Truth and Falsehood," op. cit., Burtt, "Further Technique for Inspiration-Expiration Ratios," op. cit.; C. Landis and R. Gullette, "Studies of Emotional Reactions," J. Comp. Psych., 5 (1925), 221-253; Larson, "Modification of the Marston Deception Test," op. cit.

[93] Loc. cit., Note #91.

such as, electric shock,[94] rewards,[95] loss of self esteem,[96] personally relevant material,[97] and awareness of the testing situation.[98] While many of these devices have apparently increased motivation to deceive, there is little evidence that the level of affect approaches that in real-life. Of course, it is also possible that no artificial device used in the laboratory can make the consequences of deception as real as those encountered in life. In other words, laboratory motivational-devices are ipso facto rewards for successful deception; the subject loses nothing for failing to deceive. On the other hand, real-life subjects may lose something very consequential if they fail to deceive; the liar may be subject to criminal prosecution, lose a job, etc. Likewise, the truthful person in real-life fears the consequences of being erroneously found to be a "liar"; he is highly motivated not to deceive and to do all he can to succeed in "passing" his test.

---

[94] Lykken, "The GSR in The Detection of Guilt," op. cit.

[95] Davidson, "Validity of the Guilty Knowledge Technique: The Effects of Motivation," op. cit., 62-65; Lykken, "The Validity of the Guilty Knowledge Technique: The Effects of Faking," op. cit.; Barland, "An Experimental Study of Field Techniques in Lie Detection," op. cit.

[96] L. Gustafson and M. Orne, "Effects of Heightened Motivation on the Detection of Deception," J. App. Psych., 47 (1963), 408-411.

[97] R. Thackray and M. Orne, "Effects of the Type of Stimulus Employed and the Level of Subject Awareness on the Detection of Deception," J. App. Psych., 52, 3 (1968), 234-239.

[98] Ibid.

rea

by

fir

who

tan

rat

dec

giv

wer

ind

sig

The

sup

eve

dur

lif

acc

inv

tha

Str
tic

and
spo
ter
632

Two studies which purported to assess the effects of real-life stress on laboratory lie-detection were conducted by Kugelmass and Lieblich,[99] and Kugelmass, et al.[100] In the first of these studies, card-tests given to police trainees who apparently considered their successful deception important to their future, were evaluated. Both GSR and heart rate were considered, but GSR was clearly more indicative of deception than heart rate. In the second study, card-tests given to actual criminal suspects as part of their examination, were evaluated. Again GSR-responses were clearly superior indicators of deception; heart rate responses were not significantly different from chance as deception indices. The results of these studies seem to indicate that GSR is superior to heart rate as an indicator of deception. However, it is questionable whether or not the level of affect during card-tests, even though included as a part of a real-life examination, is the same as the level of affect which accompanies personal questioning concerning possible criminal involvement. In fact, because many field examiners report that GSR is highly effective during card-tests in actual

---

[99] S. Kugelmass and I. Lieblich, "Effects of Realistic Stress and Procedural Interference in Experimental Lie Detection," J. App. Psych., 50, 3 (1966), 211-216.

[100] S. Kugelmass, I. Lieblich, A. Ben Ishai, A. Opatowski, and M. Kaplan, "Experimental Evaluation of Galvanic Skin Response and Blood Pressure Change Indices During Criminal Interrogation," J. Crim. Law, Crim., and Pol. Sci., 59 (1968), 632-635.

exa

pre

tha

typ

is

se.

Lie

an

ing

ele

cat

in

in

sta

ca

re

er

re

eq

examinations[101] and yet relatively ineffective in tests preceeding and following the card-test, it seems indicated that either a subject's level of affect varies with the type of questions asked, or that "arousal" or "attention" is more important to the success of GSR than is affect per se.

Lie-Detection Equipment

Another reason for the disparity between laboratory and field lie-detection studies concerns the type of testing-apparatus employed. Laboratory apparatus, particularly electrodermal measuring devices, are usually highly sophisticated, while field equipment is relatively simple. However, in spite of the differences in equipment used, there is increasing evidence that this does not account for any substantial difference in results. Orne has found no significant difference between the two types of equipment with respect to results obtained and laboratory studies have employed field apparatus without noticeable differences in results; electrodermal activity, regardless of the type of equipment employed, maintained superiority in lie-detection.[102]

---

[101] Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 33.

[102] M. Orne, untitled manuscript (paper presented to American Polygraph Association, Silver Springs, Maryland, 1969); see also: Barland, "An Experimental Study of Field Techniques in Lie Detection," op. cit.; and Orne, "Implications of Laboratory Research for the Detection of Deception," op. cit. wherein he expresses the belief that field GSR electrodes can be improved to increase the effectiveness of this measure (in the field), 196.

Use of Control Questions

Leading field examiners invariably employ some variation of the control-question technique in conducting lie-detection tests. Simply stated, control questions are designed to channel the psychological set of truthful subjects away from relevant questions and towards the control questions. Lying subjects, on the other hand, are presumed to be psychologically set to the relevant questions. Hence, consistently greater physiological responses to control questions are considered indicative of truthfulness regarding relevant questions, while consistently greater responses to relevant questions are suggestive of lying. The use of control-questions reportedly has significantly increased the ability of field examiners to discriminate between truthful and lying persons and at the same time has lowered the number of inconclusive tests.[103]

The fact that control-questions are generally not used in laboratory studies may be one reason that laboratory studies find cardiovascular and respiratory activity less effective in detecting lies than is electrodermal activity. For example, control questions as used in field settings are generally "worked up" with the subject to insure that the question involves personally relevant material, and that the subject will either lie or have doubts about the accuracy of

---

[103]Reid, "A Revised Questioning Technique in Lie Detection Tests," op. cit., 547.

his answer to the question.[104] In laboratory studies then, control questions could conceivably heighten a person's interest or concern for the test and possibly would result in greater differential response. The fact that personally relevant material does increase response in laboratory studies has been consistently reported,[105] and at least one laboratory study using control-questions has found that both respiration and cardiovascular activity did significantly discriminate the "liars" from the "truth-tellers".[106]

Summers[107] and Kubis,[108] both claimed accuracy-rates of over 95 percent. Significantly both of them employed "emotional standard" questions, "highly charged emotional issues selected from a study of the life history of the suspect."[109] While these "emotional standard" questions only remotely resemble control questions used today, it is clear

---

[104]G. Harman and J. Reid, "The Selection and Phrasing of Lie-Detector Test Control-Questions," J. Crim. Law, Crim. and Pol. Sci., 46 (1955), 578-582.

[105]J. Berkhout, D. Walter and W. Abey, "Autonomic Responses During a Replicable Interrogation," J. App. Psych., 54 (1970), 316-325; Thackray and Orne, "Effects of the Type of Stimulus Employed and the Level of Subject Awareness on the Detection of Deception," op. cit., 234-239.

[106]Barland, "An Experimental Study of Field Techniques in Lie Detection," op. cit.

[107]Summers, "Science can get the Confession," op. cit.

[108]J. Kubis, "Electronic Detection of Deception," Electronics, 18 (April, 1945), 192-212.

[109]J. Kubis, "Experimental and Statistical Factors in the Diagnosis of Conciously Suppressed Affective Experiences," J. Clin. Psych., 6 (1950), 13.

from

was

be c

tion

prov

This

ther

ques

arbi

or l

in t

the

or n

simp

iner

ques

In t

use

is q

prim

subj

try,

in I

from Summers' description that their function in the test was the same: to evoke reactions from a suspect which could be compared to reactions on relevant (crime-related) questions. In other words, the use of control-type questions provides a means of using each person as his own control. This is in contrast to some laboratory studies wherein there may be no real individual "control"; reactions to questions are evaluated across individuals according to some arbitrarily-assigned value; hence, all are judged truthful or lying according to the same criterion. Moreover, even in those laboratory studies which use individual "controls", the "control response" is that which occurs to irrelevant or non-critical items. That is, many laboratory researchers simply do not understand the "controls" used by field examiners;[110] they term as control-questions those kinds of questions which field-practitioners label as irrelevant. In the most recent study which attempted to approximate the use of control-questions as used by field practitioners, it is questionable if the controls were entirely adequate, primarily because they were not individually tailored to subjects.[111]

---

[110] See: Lykken, Psychology and the Lie Detector Industry, op. cit., 24-26.

[111] Barland, "An Experimental Study of Field Techniques in Lie Detection," op. cit., 40.

The

real

dete

lyin

Dece

for

wher

to c

duce

test

tena

cedu

pres

Dech

can

Berg

in d

to c

cit.
Dech

Test

of I
2 (1

The Role of Lying

Lykken has proposed that field examiners are not really in the business of lie-detection but rather guilt-detection.[112] If this is so then it seems that the act of lying per se would have little effect on field procedures. Recent evidence tends to support this hypothesis, at least for some persons.[113] The use of a "silent-answer test" wherein the person is instructed not to vocalize answers to questions and thus not really lie, has been shown to produce deception-criteria equal to and at times superior to tests which require vocal answers. Unfortunately, the maintenance of deception-responses in such a silent-answer procedure does not hold true for all persons; nor is there at present any complete understanding of the psychological mechanisms involved in such a silent-answer test.

Contradictory evidence concerning the role of lying can be found in laboratory studies. Kugelmass, Lieblich and Bergman reported that there were no significant differences in detection rates whether subjects answered "yes" or "no" to cards chosen from a deck.[114] On the other hand, Gustafson

---

[112] Lykken, "The GSR in the Detection of Guilt," op. cit., 385; Lykken, "The Validity of The Guilty Knowledge Technique: The Effects of Faking," op. cit., 258.

[113] Horvath and Reid, "The Polygraph Silent Answer Test," op. cit.

[114] S. Kugelmass, I. Lieblich, Z. Bergman, "The Role of Lying in Psychophysiological Detection," Psychophysiology, 3 (1967), 312-315.

and

car

ver

to

in

Sco

pos

fie

"ma

ced

vas

lab

hav

gat

ele

or

ins

act

of

and Orne reported that subjects answering "no" to chosen cards were detected more often than subjects giving no verbal answer; subjects required to make a word association to each question were detected less frequently than subjects in the other two groups.[115]

Scoring Response Data

In the final analysis, there is at least one other possible explanation for differences between laboratory and field lie-detection: objectively scoring response-data may "mask out" important information. Indeed, the complex procedures necessary for the objective scoring of both cardiovascular and respiratory activity have been one reason that laboratory investigators, even though recording such activity, have not evaluated it.[116] Moreover, from the evidence gathered by Kubis it is evident that visual inspection of electrodermal response-data by experienced personnel is equal or perhaps superior to objective techniques, and that visual inspection of cardiovascular, respiratory, and electrodermal activity as a unit can lead to high accuracy-rates independent of interaction between the subject and examiner.[117] This is

---

[115] L. Gustafson and M. Orne, "The Effects of Verbal Responses on the Laboratory Detection of Deception," Psychophysiology, 2 (1965), 10-13.

[116] S. Kugelmass, Effects of Three Levels of Realistic Stress on Differential Physiological Reactivities, Tech. Report, 63-61 (report prepared for Air Force Office of Scientific Research, European Office, Aerospace Research, U.S. Air Force, Hebrew University of Jerusalem, Israel, Aug., 1963).

[117] Kubis, Studies in Lie Detection: Computer Feasibility Considerations, op. cit.

consistent with the results of studies using field-obtained polygraphic records.[118]

The Reliability of Lie-Detection

The reliability of polygraphic procedures has received considerably less attention than its validity. And, of course, this is quite natural since reliability refers only to the degree of consistency of judgments between polygraphic examiners or examinations irrespective of the "correctness" of the judgments. For example, Dearman and Smith reported an instance of an individual being given independent polygraphic examinations by several different examiners, all of whom concluded that the individual had not told the truth in answering the question, "Did you steal any money from the bank or its customers?" In other words, in this instance the reliability of the examiners' judgments was perfect. However, Dearman and Smith pointed out that in their judgment, based on psychiatric evaluations, the individual in question had told the truth to the test question; in other words, while the reliability between the examiners was high, validity, according to Dearman and Smith's interpretation, was low.[119] This example, of course, concerns the reliability

---

[118] See: Holmes, "The Degree of Objectivity in Chart Interpretation," op. cit.; Horvath and Reid, "The Reliability of Polygraph Examiners Diagnosis of Truth and Deception," op. cit.; Hunter and Ash, "The Accuracy and Consistency of Polygraph Examiner's Diagnosis," op. cit.; S. Hathaway and C. Hanscom, "The Statistical Evaluation of Polygraph Records," Academy Lectures on Lie Detection, II, V. Leonard (Ed.) (Springfield, Ill.: C.C Thomas, 1958), 118-136.

[119] Dearman and Smith, "Unconcious Motivation and the Polygraph Test," op. cit.

of

bee

fie

agr

rec

men

lat

sho

iss

ess

Lab

by

90

sit

act

use

eve

cie

reg

two

que

pe

in

ta

of the complete polygraphic procedure; as such it has not been adequately reported in the literature. The reported field reliability studies deal rather with the degree of agreement between evaluators when judging the same polygraphic recordings, or with the consistency of one evaluator's judgment of the same recording two or more times. It is these latter studies which will be discussed here shortly; it should be noted that many of them deal indirectly with the issue of validity, although such a consideration is not essential for reliability-studies.

Laboratory Studies

The earliest of the reliability-studies was reported by Rouke in 1941. Two groups of subjects, 80 delinquent and 90 non-delinquent boys, were tested in an "experimental situation designed to simulate closely the elements in the actual investigation of criminal cases."[120] The tests given used only a psychogalvanic (GSR) measure. There was, however, a very close correspondence (C, contingency coefficient, = .72) between the ratings (evaluations) of the same records (tests) by the same evaluator at different times, and two judges independently reviewing the records of the delinquent and non-delinquent boys agreed in their judgments 88 percent and 91 percent of the time, respectively.

---

[120] F. Rouke, "Evaluation of the Indices of Deception in the Psychogalvanic Technique" (unpublished Ph.D. dissertation, Fordham University, 1941), 80.

repo

expe

deta

Firs

tha

carc

eva

whor

the

Kub

the

to

sub

int

was

by

qui

int

tha

jud

ner

bil

The most thorough study of reliability to date was reported by Kubis who conducted an elaborate series of experiments on lie-detection. While it is not necessary to detail them here, there are several points of interest. First, recordings were obtained by means of a polygraph; that is, respiration, electrodermal activity (GSR), and cardiovascular activity were recorded. Second, the examiner-evaluators used by Kubis were trained psychologists, all of whom were given a special "three-month training course in the theory and practice of 'lie detection'".[121] Third, Kubis was able to assess the reliability with which each of the physiological measurements was interpreted and was able to compare the reliability of examiners who interacted with subjects to that of evaluators who had not engaged in such interaction.

In Kubis' study each of the polygraphic recordings was evaluated by the examiner who had done the testing, and by two independent evaluators. While all evaluations were quite accurate the reliability of the judgments is of major interest here. Kubis found in one section of his experiment that there was an average 78 percent agreement between the judgments made by examiners and independent evaluators; judgments made by only independent evaluators agreed, on the

---

[121]Kubis, Studies in Lie Detection: Computer Feasibility Considerations, op. cit., 28.

average, 81 percent of the time.[122] Similar results, ranging from 72 percent to 87 percent were reported in another section of Kubis' experiments.[123]

It should be noted that the reliability reported by Kubis varied with the particular physiological parameter evaluated, GSR being judged more reliably than either respiration or cardiovascular recordings. Similar results have been reported by Barland who submitted experimentally-derived polygraphic recordings to a group of independent evaluators, all trained polygraph examiners.[124]

Kubis also reported that independent evaluators had "greater confidence in those decisions which were ultimately verified as correct than they did in those which were incorrect."[125] Moroney, using an experimental lie-detection situation but recording only GSR, substantiated Kubis' results: the more confident evaluators were in their decisions, the more likely they were to be correct; that is, the more ambiguous the recordings, the greater the likelihood of error.[126]

---

[122] Ibid., 44.

[123] Ibid., 48.

[124] Barland, "The Reliability of Polygraph Chart Evaluations," op. cit.

[125] Kubis, Studies in Lie Detection: Computer Feasibility Considerations, op. cit., 68.

[126] W. Moroney, "The Detection of Deception as a Function of PGR Methodology" (unpublished Ph.D. dissertation, St. Johns University, 1968, Ann Arbor, Mich.: University Microfilms, 1969, No. 69-7125).

In a recent study Barland submitted the polygraphic recordings of 72 subjects involved in a hypothetical crime to a group of five independent evaluators, all experienced polygraphic examiners. Rather than having the evaluators make dichotomous or trichotomous (i.e., "guilty", "innocent", or "inconclusive") judgments, he asked them to evaluate the recordings in accordance with the numerical scoring-system developed by Backster.[127] Hence, a total numerical score was obtained for each subject's records (tests) from each of the evaluators. By considering evaluators in pairs, and including his own evaluations, correlations (Pearson product-moment) between all possible pairs of evaluators were computed; such correlations ranged from .78 to .95 with a mean of .86, indicating a very high reliability among the evaluators. Said another way, Barland found that out of 559 instances of two examiners arriving at a definite judgment of truth or deception, agreement occurred 534 times, or 95.5 percent of the time.[128]

Other investigators have also reported high reliability in the evaluations of physiological data gathered in experimental lie-detection settings. Van Buskirk and

---

[127]See pages 37-39.

[128]Barland, "The Reliability of Polygraph Chart Evaluations," op. cit., 5.

Marcuse, for example, using standard field polygraphic equipment and the card-test, had two evaluators judge the same 50 records at two different times one month apart. "The results indicated 84 percent agreement on cards and 94 percent agreement on records between these two judgments."[129] Bitterman and Marcuse reported that their judgments concerning the classification of response-data in cardiovascular tracings were highly reliable (C = .96 and .92); a third classification by an independent evaluator of the recordings demonstrated that the authors' classification was substantially reproducible.[130] And, in a study reported by Heckel, et al., a hypothetical crime was set up in such a way that three groups of five subjects each were led to believe that they were suspected of stealing money from the experimenter's wallet. One group consisted of "normal" males recruited from a local educational institution; the other two groups consisted of males under phychiatric care and diagnosed as either "non-delusional" (psychoneurotics) or "delusional" (psychotics). Although none of the subjects were, in fact, guilty of the theft, they were all given polygraphic tests by a skilled examiner; the purpose of giving such tests was to determine if physiological reactions to the testing differed between the groups, affecting the interpretation of recordings.

---

[129] Van Buskirk and Marcuse, "The Nature of Errors in Experimental Lie Detection," op. cit., 188.

[130] M. Bitterman and F. Marcuse, "Cardiovascular Responses of Innocent Persons to Criminal Interrogation," Amer. J. Psych., 60 (1947), 407-412.

Following the administration of all polygraphic tests, the recordings were submitted to a group of four trained examiners asked to judge if the recordings indicated deception or no deception, or were inconclusive. Complete agreement on the control-subjects prevailed between the four evaluators, and, in general, reliability decreased for the "psychiatric" subjects although "overall reliability of ratings was quite high."[131] This suggests that polygraphic recordings of persons indicating psychiatric maladjustment may be subject to erroneous judgments, i.e., less valid, and that examiners' agreement on recordings obtained from such persons may be less than the recordings from "normal" persons.

It is important to note that all of the above studies dealt with data derived from experimental lie-detection situations. It is generally agreed that such data are not necessarily related to those obtained in field situations. Therefore, we must turn to an analysis of field studies which have looked at the issue of reliability.

Field Studies

In a recent study, Horvath and Reid submitted the polygraphic recordings of forty subjects, 20 verified truth-tellers and 20 verified liars, along with brief factual information of the investigations in which the subjects were

---

[131] R. Heckel, J. Brokaw, H. Salzberg and S. Wiggins, "Polygraphic Variations in Reactivity Between Delusional, Non-Delusional and Control Groups in a 'Crime' Situation," J. Crim. Law, Crim. and Pol. Sci., 53 (1962), 382.

involved, to a group of ten examiner-evaluators. The evaluators were asked to identify the truth-tellers and liars. In spite of their minimal information about the investigations, they were able to achieve an average rate of agreement of 87.8 percent, the more experienced 91 percent, the less 79 percent. It is noteworthy that the evaluators in this study were deliberately given polygraphic recordings felt by the authors to be difficult to interpret, that is records not dramatically indicative of truth-telling or lying.[132]

Hunter and Ash have reported the results of a study which essentially dealt with test-retest reliability. The polygraphic records of ten verified truth-tellers and ten verified liars were given to a group of seven examiner-evaluators at two different times; a minimum of three months elapsed between the two evaluations, no evaluator being told that he would be dealing with the same polygraphic records on both occassions.

The results of the Hunter and Ash study were quite similar to those reported by Horvath and Reid, even though the evaluators and the polygraphic records were different. The evaluators achieved an average accuracy of 86 percent in correctly identifying the truthful and deceptive subjects, the range was between 82.5 and 90 percent. Moreover, the reliability between initial and subsequent evaluations was

---

[132]Horvath and Reid, "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," op. cit., 278.

quite high, 85 percent ranging from 75 to 90 percent. However, unlike Horvath and Reid, who found that errors seemed to favor the lying subject, Hunter and Ash reported that errors were almost identically balanced, that is, "false positives" (reporting a truth-teller as a liar) were made as often as "false negatives".[133]

The Horvath/Reid and Hunter/Ash studies appear to deal with the issue of validity. In a sense they do; however, they should not be viewed as providing direct evidence of the validity of field lie-detection procedures. This is primarily because in these studies the polygraphic records evaluated were selected from cases where the testing examiner correctly identified the guilty person. It can be argued that in such cases the non-polygraphic sources of information available to the examiner aided considerably in conducting the examination; better factual information might have allowed him to formulate more appropriate test questions (affecting the response data on the recordings); or to vary his pre-test interview in a way that made it possible to obtain more suitable recordings than would otherwise be obtained. In other words, while these studies suggest that blind analysis of physiological data can lead to considerable accuracy, the chief value of the studies is as reliability assessments; that is, independent evaluators trained in the same tradition, can

---

[133]Hunter and Ash, "The Accuracy and Consistency of Polygraph Examiner's Diagnoses," op. cit., 372.

consistently identify those physiological changes believed to be associated with deception.

Discussion

Most research dealing with "lie-detection" has been done in the laboratory. Unfortunately, such research, while important for understanding the mechanisms which underlie detection of deception, is not necessarily applicable to real life. For example, laboratory researchers almost without exception report that electrodermal activity is the most valid and reliable indicator of deception; field practitioners, on the other hand, claim that for their purposes other physiological measures are more useful. Nor are the types of testing most often used in the laboratory - relevant-irrelevant tests - believed to be adequate in the field where control-question tests predominate. While it is unlikely that the reasons for these and other differences between laboratory and field lie-detection will be easily and quickly resolved, there are some approaches to these issues which provide suggestive evidence.

Studies such as Bersh's using completely independent criteria to validate field polygraph examiners' judgments, seem to hold promise. Also, studies which require independent evaluators (trained polygraphic examiners) to make judgments of truth and deception solely on the basis of physiological data obtained in the field appear to be useful. Unfortunately, reported studies in the latter category raise as many questions as they answer.

The major deficiencies in those studies requiring independent judgments to be made on polygraphic recordings obtained in field settings are these:

1) Except for the study of Holmes whose data were inadequate for making reliability-assessments, none of these studies have dealt with judgments made by polygraph examiners employed by law-enforcement agencies. Horvath and Reid and Hunter and Ash, for example, evaluated the judgments of examiners employed by a private agency; these examiners were more highly educated and had received more training in the polygraphic technique than most police polygraphic examiners.[134] It is not unreasonable to suspect that such education and training influenced the results. Thus, it is important that such studies be replicated with examiners who are representative of those employed by police agencies and as such more likely to deal with persons whose liberty may depend on the outcome of the polygraphic examination. Moreover, in judicial proceedings the police examiner is more likely to be called upon to testify as to the results of polygraphic examinations, assuming that such evidence becomes generally admissable for such purposes in the future.

---

[134] Horvath and Reid and Hunter and Ash evaluated judgments of examiners who by state law were required to possess at least a Baccalaureate degree and to undergo a training program of six months duration. At the present time such minimal qualifications are not required of polygraphic examiners employed in other jurisdictions. For a discussion of this topic, see: C. Romig, "The Status of Polygraph Legislation of the Fifty States," Police, 16, No. 1, 2, 3 (1971), 35-41, 54-61, 55-61, respectively.

2) All of the field studies, as well as many of the laboratory studies, have used polygraphic records obtained from persons tested by only one examiner. For instance, Horvath and Reid and Hunter and Ash employed recordings in each instance originally obtained by one of the authors. Obviously, the use of such records at least partially controls for the nature of the interaction between the subject and the examiner, interaction believed to have an affect on the nature of the recordings obtained. Hence, these studies show only that when this interaction is controlled in such a manner, other examiners trained within the same tradition can make independent judgments which are "accurate" and reliable. Whether or not similar results would obtain if the interaction were not accounted for in the manner described above is not known.

3) The recordings used in the reported field studies do not necessarily constitute a representative sample of any pre-defined population. Horvath and Reid reported results obtained when evaluators judged recordings selected because they were believed to require skill to interpret. Selection of recordings in such a manner makes it difficult to draw any valid general conclusions from results. Moreover, all of the studies of recordings obtained from field situations used those ultimately verified as being of either a "truth-teller" or "liar", according to corroborated confessions. Such criteria, of course are necessary for estimating

accuracy, but are not required for assessments of reliability. And by using only recordings "verified" in such a way, generalizations are seriously restricted since the majority of all persons tested by polygraphic examiners are not verified as truth-tellers or liars by confessions or any other information.[135]

It has been suggested that persons presumed prior to testing to be liars undergo an examination somewhat different from those presumed to be truth-tellers. By similar reasoning persons involved in investigations which are eventually verified by someone's confession, would undergo examinations differing from those not so verified; factual information, behavioral characteristics, etc. would provide more, or "better", clues to the examiner in the verified investigations, or perhaps, the resulting polygraphic records, for some reason, would be of a better quality. In other words, it is important to assess accuracy and reliability in terms of records obtained from both verified and unverified investigations, using in the latter instance the testing examiner's judgment for comparison with independently made judgments. As Holmes,[136] and others[137] have

---

[135] See: Inbau and Reid, Lie Detection and Criminal Interrogation, op. cit., 110-113.

[136] Holmes, "The Degree of Objectivity in Chart Interpretation," op. cit.

[137] Horvath and Reid, "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," op. cit.; Hunter and Ash, "The Accuracy and Consistency of Polygraph Examiner's Diagnoses," op. cit.

pointed out, there is some reason to believe that the judgment of the testing examiner, because it includes an evaluation of non-polygraphic sources of information, is more apt to be correct than an evaluation based on polygraphic records alone.

4) Some of the previous research suggests that experience in giving polygraphic examinations affects the reliability of judgments of truth and deception. Horvath and Reid, for instance, compared judgments of examiners with less than six months' experience and still undergoing training, to those of examiners with more than six months' experience; the former group were less reliable than the latter. Hunter and Ash reported similar results using a different group of examiners and, moreover, found that the examiner with the most experience was more consistent in his judgments than all other examiners. In further recognition of experience as an important determinant of the ability of an examiner to interpret polygraphic records is the proposal by Reid and Inbau that examiners selected for giving testimony in a courtroom should have more than five years experience in field testing.[138] While there are many ramifications to such a proposal, the implication that experience is an important determinant of success is clear.

---

[138]Reid and Inbau, Truth and Deception; The Polygraph ("Lie Detector") Technique, op. cit., 257.

One aspect of the difference in results between experience levels as discussed above is easily accounted for: the inexperienced examiners had not yet completed their training. It is still not known if fully trained and experienced examiners will be more "accurate" and consistent in their judgments than those with less experience, although the Hunter and Ash results suggest that this is so.

5) Another shortcoming of the previous studies is that the nature of the investigation from which polygraphic records were obtained was not controlled. Do differences in results depend upon the nature of the crime? For instance, if recordings were drawn from investigations classified according to crimes against a person or property crimes,[139] it seems reasonable to suspect that in the former category a testing examiner would have more factual information at his disposal than in the latter. Offenses such as rape or armed robbery involve a victim who is usually capable of identifying a suspect or, at least, of relating precise details regarding how and where the offense occurred. Even in homicide cases, where naturally a victim is incapable of providing details, the details possible seem to be generally quite adequate, such offenses usually giving the

---

[139] Such classification based on the presumed nature of involvement of the victim; direct involvement, such as in rape, murder, armed robbery, assault, and indecent (sexual) liberties, leading to classification as "crimes against a person"; less apparent involvement of the victim, such as breaking and entering, arson, and larceny, leading to classification as "property crimes".

high

as b

not

rela

tion

pers

ther

dece

more

invo

to

rec

mor

pro

of

bas

ine

phy

per

may

highest police-clearance rate.[140] In property crimes such as burglary, larceny, and arson, on the other hand, usually not directly involving either a victim or witness capable of relating precise details about the offense, factual information is less apparent.

The advantages which an examiner may have in testing persons suspected of committing a crime against a person, then, could conceivably influence his judgment of truth and deception. That is, the examiner might be inclined to give more credence to factual information when testing persons involved in crimes against a person, and to give less weight to the physiological responses observed in polygraphic records; or, assuming that crimes against a person involve more detailed information for an examiner's use, it seems probable that such information profoundly affects the outcome of the examination; detailed information provides a firmer basis for formulating test questions, which, as field examiners are well aware,[141] is an important determinant of physiological responsitivity. Also, as Orne has suggested, persons who examiners prior to testing presume to be liars may undergo an examination somewhat different from those

---

[140] Federal Bureau of Investigation, Uniform Crime Reports for the United States: 1972 (Washington: Government Printing Office, 1973), 115.

[141] Reid and Inbau, Truth and Deception, The Polygraph ("Lie Detector") Technique, op. cit., 16-21; R. Arther, "Crime Question Wording," J. Polygraph Studies, 4 (Sept.-Oct., 1969), 1-4.

presumed to be telling the truth,[142] the examiner's bias influencing the nature of the questions and therefore the responses as polygraphically recorded. Rosenthal's work too makes it difficult to believe that such bias does not exist in polygraphic testing.[143] It seems reasonable then to conclude that bias is more likely when an examiner has access to detailed factual information which may strongly implicate or exculpate a person in involvement in a criminal offense. And, as has already been suggested, such detailed information seems more available in investigations involving crimes against a person than in those involving property crimes.

Summary

In this chapter the literature pertaining to the procedures, validity, and reliability of lie-detection in both field and laboratory settings was discussed. The procedures used in the field-setting make lie-detection there akin to a diagnostic technique whose efficacy is determined by the interaction of examiner and subject as well as by polygraphic recordings. In contrast, laboratory procedures are rarely affected by such interaction. Rather, polygraphic recordings

---

[142] Orne, "Implications of Laboratory Research for the Detection of Deception," op. cit., 175-177.

[143] R. Rosenthal, Experimenter Effects in Behavioral Research (New York: Appleton-Century-Crofts, 1966).

alone, i.e., physiological measurements made during a series of tests, which also differ in nature from those used in the field, constitute laboratory lie-detection.

Because of the numerous and significant differences between laboratory and field procedures and goals, it is, in general, misleading to apply the results of laboratory research to the typical field situation. In spite of this difficulty, however, there is substantial agreement that lie-detection is a relatively valid and reliable method of determining truthfulness and deception; that is, judgments based upon lie-detection tests are correct too often to be considered coincidental, and the physiological responses thus measured and recorded provide a basis for substantial replication of judgments made on them.

Many field-practitioners of lie-detection claim that relatively recent developments in administering such tests, e.g., the control-question procedure as well as the standardization of procedures between examiners, provide an adequate basis for conducting meaningful research on field-gathered data. And some research recently reported suggests that this is indeed true, despite the many important questions still unanswered. There remains a need to replicate this research and to introduce innovations to clarify and supplement the findings so far reported in the literature.

Chapter III

METHOD

In this chapter the characteristics of the sampling procedure, the polygraphic records used, the evaluators, the operational measures, hypotheses, and statistical analysis and design will be presented. First, however, a discussion of certain general characteristics of the study is in order; the source of the data as well as the nature of the testing procedure and apparatus employed will be considered.

General Considerations

Source of Polygraphic Data

A large state police department (SPD) located in the mid-western states provided the researcher access to its files containing data pretaining to polygraphic examinations conducted by employees of this agency. While the SPD had such files at ten locations or posts throughout the state, it was decided that only those files at a post located in a large metropolitan area would be used for this study. There were two major reasons for this choice. First, it was believed that the method of filing data at this post would facilitate sampling procedures. And, second, the examinations conducted there generally involve a wider variety of serious

criminal offenses than at other posts, which, for purposes of the study, was desirable.

At the site selected, data are compiled and filed in the following manner. When a complaint of criminal conduct comes to the attention of the law-enforcement agency[1] the person against whom the complaint is made is asked to undergo polygraphic examination. If he agrees, the data pertaining to this examination is placed in a case folder, on the outside of which are written the person's name, the nature of the investigation (homicide, rape, etc.), other identifying data (e.g., complaint number) and the outcome of the examination. If another person is given an examination with respect to the same complaint, data pertaining to that examination are added to the same folder as are appropriate notations on the outside. Hence, a common folder contains all data pertaining to examinations relevant to the same complaint, the outside of such folders indicating the nature of the contents.

Examination Procedure

All polygraphic records used in this study were obtained from examinations conducted by employees of the SPD. These examiners had all received their initial training

[1]The SPD conducts polygraphic examinations not only for its own investigations but also for other law enforcement agencies making appropriate requests.

at

Am

th

fo

te

th

se

a

ho

as

of

in

su

et

ex

in

Cit

Ac

to

Ac

In

at a nationally recognized training school[2] certified by the American Polygraph Association.[3]

All examinations were conducted in accordance with the Control Question Technique noted in Chapter II. The following discussion further describes certain aspects of this technique in greater detail, highlighting differences between the procedures employed by the SPD examiners and those presented elsewhere.

Pre-test interview.--The SPD examinations consist of a pre-test interview and polygraphic testing. The interview, however, is essentially an eclectic one, combining certain aspects of the interview procedures advocated by proponents of the various approaches to CQ-testing. For instance, during the interview the examiner discusses in depth with the subject, the subject's likes and dislikes, hobbies, education, etc. Such a discussion is similar to that used by military examiners as reported by Barland and Raskin.[4] Also included in the interview are questions specifically designed to elicit

---

[2]National Training Center of Lie Detection, New York City, New York.

[3]N. Ansley (Ed.), "A.P.A. Accepted Polygraph Schools," American Polygraph Association Newsletter (December/January, 1974), 14.

[4]G. Barland and D. Raskin, "The Use of Electrodermal Activity in the Detection of Deception," pre-publication copy to appear in: W. Prokasy and D. Raskin (Eds.), Electrodermal Activity in Psychological Research (New York: Academic Press, in press), 5-8.

behavioral cues from the subject, questions "borrowed" from the interview procedure used by proponents of the Reid technique.[5] The interview ends with a procedure advocated by Arther, an extended explanation of the polygraph instrument and the nature of "lie detection", etc.[6]

In spite of the eclectic nature of this SPD interviewing, the procedure is consistent with CQ testing: there is no intensive or accusatory questioning prior to (or during) polygraphic testing, and all test questions are reviewed exactly as they will be asked during actual testing; the questions are worded or phrased in such a way that the subject is certain that he understands them and that he can answer them with either a "yes" or a "no".

Polygraphic testing.--The interview, which usually lasts between 45 and 90 minutes, is followed by the polygraphic testing. The polygraphic attachments are placed on the subject, the pneumograph (respiration recording) and the GSR units are activated and recordings made for about a minute in order to assure that they are suitable. The subject is told that the testing is about to begin and is reminded to

---

[5]See: J. Reid and F. Inbau, Truth and Deception, The Polygraph ("Lie Detector") Technique (Baltimore: Williams and Wilkins, 1966), 10-16; F. Horvath, "Verbal and Nonverbal Clues to Truth and Deception During Polygraph Examinations," J. Pol. Sci. and Adm., 1 (1973), 138-152.

[6]R. Arther, "The Heart and You" (unpublished, undated manuscript, National Training Center of Lie Detection, New York).

answer all questions either "yes" or "no", as he did during the run-through. The cardio-cuff is then inflated and the questioning begins.

The test questions are asked at about 15-20 second intervals in a pre-determined order. During the questioning the examiner marks the following on the chart paper: the sensitivity setting of the GSR amplifier, the pressure in the cardio-cuff, the points at which each question starts and ends, the number of each question, and the subject's answers. Any adjustments to the tracings made by the examiner or "artifacts" caused by the subject's movements, etc. are also appropriately noted. When each test question has been asked once, and the test concluded the pressure in the cardio-cuff is again noted and the cuff, deflated. The pneumograph and GSR units remain in operation for a short period following deflation of the cuff.

A test usually lasts for about three minutes, after which the examiner notes the subject's name, the date, his own initials, and the number of the test in the sequence. This is done on the chart paper at a point prior to where the cardio-cuff was inflated at the beginning of the test. As explained in Chapter II, however, a battery of such tests is conducted with each subject before the examination is completed.

The basic battery of tests used by the SPD examiners consists of at least CQ Test #1, a "card" or "number" test

and then a third test, a repetition of Test #1. In some instances, only these three tests are conducted; in others, the examiner may conduct additional tests, making use of various stimulation strategies.[7] Such additional tests almost always include a "mixed question" test as the fourth in the series, although in rare instances the fourth may also be a "yes" test, or a "yes-no" test.[8] Regardless of which test follows the basic battery, however, additional tests are always consecutively numbered and the nature of the stimulation strategy used by the examiner is indicated on the chart paper according to standardized notation. In other words, it is possible on review of any subject's records (tests) to determine where in the sequence a test was conducted as well as the nature of the test itself.

Sequencing of test questions.--As explained above, the sequencing of the questions in a given test is predetermined, and consistent with the variation of CQ testing used by the SPD examiners.[9] Perhaps, the sequence can be best explained by discussion of an example. Assume a burglary has taken place, a polygraphic instrument stolen. Questions considered pertinent in such a case would be as follows:

---

[7]See Chapter II, pages 29-32.

[8]R. Golden, "The Yes-No Technique" (paper presented at the American Polygraph Association Seminar, August 1969, Houston, Texas).

[9]The SPD examiners sequence questions in a manner advocated by Arther. See: R. Arther, "Irrelevant Questions," J. Polygraph Studies, 3 (May-June, 1969), 3-4.

Position in Sequence	Numerical Designation on Charts	Type of Question	Example
1	1	Irrelevant	"Are you in Michigan now?"
2	3T	"Known Truth"	"Did you sell that polygraph to (a fictitious person)?"
3	3K	Relevant-Guilty Knowledge	"Do you know for sure who stole that polygraph?"
4	5	Relevant-Crime Related	"Did you steal that polygraph?"
5	6	Control	"Did you ever steal anything?"
6	8	Relevant-Crime Related	"Do you know where that missing polygraph is now?"
7	8GC	"Guilt Complex"-fictitious Crime	"Did you steal (fictitious item) from (fictitious person or place)?"
8	9	Relevant-Crime Related	"Did you break into (building or location from which the polygraph was stolen)?"
9	10	Control	"Did you ever lie about anything important ?"
10	11	Relevant-Crime Related	"Did you tell the complete truth about that missing polygraph?"

While other publications discuss in detail the rationale and purpose of such a sequence of questions,[10] of significance here are several points in particular. (1) The sequence, although pre-determined is not inflexible; that is, the specific nature of the investigation determines the precise wording of the questions and the elimination of certain question types. A guilt-complex question, for instance, if not useful in certain types of investigations,[11] would be replaced by a different question in the seventh position in the sequence. (2) Two control questions, each individually prepared with each subject, are always imbedded in the series. Actually, the known-truth and guilt complex questions, when asked, serve as quasi-control questions; responses to them permit estimation of a subject's response to relevant test questions when he is telling the truth (to the relevant questions). (3) The sequencing of the questions in CQ test #2 following the card test, is identical with that in CQ test #1. (4) Additional irrelevant questions, pre-reviewed with the subject, can be inserted in the sequence at the examiner's discretion; such questions would be designated on the charts as #2, #3, or #7. Finally, the designation of

---

[10] R. Arther, "Crime Question Wording," J. Polygraph Studies, 4 (September-October, 1969), 1-4; R. Arther, "Covering Two Crimes in One Examination," J. Polygraph Studies, 4 (May-June, 1970), 3-4.

[11] R. Arther, "Irrelevant Questions," J. Polygraph Studies, 3 (May-June, 1969), 3-4.

questions on the charts or records is standardized; the number 3T, for instance, always indicates that a known-truth question was asked; the numbers 6 and 10 always refer to control questions, etc. Such standardized notation facilitates one examiner's review of another's polygraphic records and allows a determination of the nature or type of questions asked.

Polygraphic apparatus.--The recording instruments used by the SPD examiners in conducting polygraphic examinations were standard field equipment made by the two major manufacturers,[12] recording respiration, cardiovascular activity, and GSR. Between the years considered in this study, 1969-1972, however, a change in instrumentation used by the SPD examiners was made; dual pneumograph units, by which both abdominal and thoracic breathing patterns could be recorded simultaneously, were added.

Sampling Considerations

Population

The case folders pertaining to all polygraphic examinations conducted at the aforementioned SPD post during the years 1969-1972, inclusive, were reviewed; eliminated were those investigations involving violations of narcotic

---

[12] During the years from which the sample was drawn the polygraphic instruments used by the SPD were manufactured by either the Stoelting Company, 424 N. Homan Ave., Chicago; or Associated Research, Inc., 3758 W. Belmont, Chicago.

laws, traffic laws, and certain other violations not readily classifiable as crimes against a person or property crimes (e.g., drunkenness). The remaining 1446 folders were used as the population from which a stratified random sample of folders was drawn.[13]

Procedure

Sampling was carried out in essentially two stages. The first stage consisted of assigning case folders to, and randomly drawing sub-samples from, eight categories according to a pre-determined stratification scheme. The stratification matrix shown in Figure 3.1 exemplifies this scheme, folders categorized as data pertaining to either verified or unverified investigations, truthful or deceptive subjects, crimes against a person or property. A verified investigation was defined as one in which a subject made a complete confession, e.g., if 10 persons were given polygraphic examinations as part of a homicide investigation, and subsequent to all examinations the tenth person made a complete confession, the investigation (folder) was considered verified.[14]

---

[13]The number of folders assigned to each stratification level in the population is given in Appendix A.

[14]In some instances notations made on the folders categorized as "verified" also indicated that the deceptive subject had either plead or been found guilty by judicial proceedings. Such notations were possible because of an informal "follow-up" procedure practiced by the SPD examiners.

An unverified investigation was defined as one in which no confession was made but in which the examiner issued a written report stating that the subject either was or was not truthfully answering questions concerning the issue at hand. "Truthful" and "deceptive" were defined by the outcome of polygraphic examinations. Crimes against a person were those with direct victim involvement, e.g., in homicide, assault, armed robbery, rape, and certain other sexual offenses; property crimes were arson, burglary, larceny, forgery, embezzlement, and malicious destruction of property.

Categories of Folder Assignment

Verified			
Truthful		Deceptive	
Crimes Against a Person	Property Crimes	Crimes Against a Person	Property Crimes
Unverified			
Truthful		Deceptive	
Crimes Against a Person	Property Crimes	Crimes Against a Person	Property Crimes

Figure 3.1.--Stratification Matrix

As is apparent from the previous discussion regarding the nature of case-folders, some folders contained data pertaining to more than one subject. This presented a problem of assignment to categories. For instance, consider the investigation of a rape case where both the victim and a

suspect are given polygraphic examinations. The victim is found to be truthful; the suspect is deceptive and subsequent to his examination confesses his guilt. It is obvious that such a folder could have been assigned to the category "verified-truthful-crime against a person" (the victim) or "verified-deceptive-crime against a person." When such a problem was encountered the folder was assigned to the "verified-truthful" category irrespective of other data included in the folder. On the other hand, if a folder contained data pertaining to a "verified-deceptive" subject and did not include "verified truth-teller" data it was, of course, assigned to the former category.

In instances of unverified examinations, folders were assigned to categories according to predominating data. If, for example, a folder contained the data of three subjects, two of whom were reported truthful and one deceptive, it was assigned to the "unverified truth-teller" category, and depending on the nature of the investigation, to either the "crime against a person" or "property crime" classification. If outcomes were balanced, a "coin toss" resolved the assignment problem.

Following the assignment procedure discussed above, all folders of each category were consecutively numbered. Then, according to a table of random numbers[15] a sample of

---

[15]L. Chao, Statistics: Methods and Analyses (New York: McGraw-Hill, 1969), 471-476.

112 folders, 14 from each category, was drawn. The sample, however, in spite of the assignment procedures mentioned above, still included some folders which contained data pertaining to subjects who fell into the same category. In the instance of the homicide previously mentioned, the case folder would have been assigned as "verified-truthful-crime-against-a-person" and would have contained the polygraphic records of the nine subjects so classified. Hence, a second stage in sampling was required.

The second stage in sampling consisted of a coin-toss decision of which subject's records should be drawn from a folder when two or more subjects fell into the same category. The purpose of the procedure was to prevent possible inclusion of records of more than one subject from each investigation. By such a restriction the records themselves were insured as independent of each other as possible. For example, if more than one subject's records were drawn from the same folder, the examiner could be reasonably assumed influenced by his knowledge of the outcome of the examination of the first subject when testing the second subject; insights gained while testing the first subject would affect the testing of the second.

Sample

The sample, then, consisted of the complete battery of polygraphic tests (record sets) of 112 persons (subjects)

involved in separate criminal investigations. The record sets of fifty-six of the subjects were verified, that is, the truthfulness of these subjects' responses (answers) to the relevant test questions was "known". An additional 56 record sets were drawn from subjects whose truthfulness was not "known" but whose responses had been designated in examiners' written reports as truthful or not.

Within each of the two major categories (verified-unverified) one-half (28) of the record sets were those of persons considered NDI (no deception indicated to relevant questions); and one-half (28) DI (deception indicated). Further, one-half (14) of the record sets within each of the NDI-DI groupings pertained to property crimes, and one-half (14) to crimes against a person.

Criteria for Record Sets

All record sets drawn in the initial sampling were reviewed by the researcher[16] and the Chief Polygraph Examiner[17] of the SPD before final selection. The purpose of the review was to insure that all record sets (or tests for each subject) met the following criteria:

1) Physiological data recorded for each subject during each test in respiration, GSR, and cardiovascular activity.

---

[16]The researcher had over six years of experience as a practicing polygraph examiner.

[17]The Chief Examiner did not serve as an evaluator in this study.

2) At least two separate control-question tests in which the relevant, irrelevant, and control questions were asked at least once per test. In addition, a standard stimulation test, commonly called a "number" or "card" test administered between the aforementioned control-question tests.

3) Records substantially free of "artifacts" such as those resulting from the subject's effort to "beat" the polygraph;[18] exception to this criterion only when such "artifacts" were apparent during a subject's "yes" test but not other tests.

4) All relevant test questions pertinent to the same specific criminal offense, i.e., burglary, rape, etc.

Mutual agreement of the Chief Examiner of the SPD and the researcher was required for retention of each record set in the sample. In the few instances when such agreement was not possible the records of another subject were substituted in accordance with the sampling procedure discussed earlier, until the sample quota was met.

Characteristics of Subjects

A summary of the background characteristics of the subjects from whom the final sample of records was obtained is displayed in Table 3.1. Further summarization of these data indicates that 92 of the subjects were Caucasian, 20 Negroid; 98 male and 14 female. The age of the subjects

---

[18] See: Reid and Inbau, Truth and Deception, The Polygraph ("Lie Detector") Technique, op. cit., 163-165, 207-218.

Table 3.1.--Background Characteristics of Subjects.

Background Characteristics	Subject Categories							
	Verified				Unverified			
	Truthful		Deceptive		Truthful		Deceptive	
	Crimes Against A Person	Property Crimes	Crimes Against A Person	Property Crimes	Crimes Against A Person	Property Crimes	Crimes Against A Person	Property Crimes
Race								
No. caucasian	13	12	11	10	12	13	10	11
No. negroid	1	2	3	4	2	1	4	3
Sex								
No. male	5	11	14	14	14	12	14	14
No. female	9	3	0	0	0	2	0	0
Age								
Range	16-45	14-50	15-57	15-30	17-49	13-67	17-47	15-54
Mean	23.1	22.6	26.6	21.1	24.0	24.9	29.2	26.5
Stan. Dev.	7.7	8.8	11.5	5.0	8.1	13.9	9.4	10.1
Education (yrs. completed)								
Range	6-12	9-12	2-12	7-12	10-12	5-12	3-12	8-12
Mean	10.6	10.9	9.8	10.3	11.6	10.4	9.3	10.8
Stan. Dev.	1.6	1.0	2.9	1.7	0.6	1.9	2.5	1.0

ranged from 13-67 with a mean of 24.8; the years of formal schooling completed ranged from 2-12 with a mean of 10.4.

Characteristics of Record Sets

The criteria cited previously were minimal, some of the record sets, due to the procedure used by the SPD examiners containing more tests than others. Moreover, due to the instrumentation change during the years from which the sample was drawn, some record sets reflected the use of a dual respiration-tracing, while others contained only one tracing. A breakdown of these differences in the record sets is given in Table 3.2, which shows that 64 of the record sets contained only the basic battery of tests (CQ Test #1, the "card" Test, and CQ Test #2) while 48 contained additional CQ tests such as the "mixed question test". A dual respiration-tracing was evident in 26 of the record sets. And, although the "yes" tests were eliminated from all record sets for reasons which will be explained shortly, it is clear from the data displayed in Table 3.2 that such tests were administered predominately to "deceptive" subjects.

Procedure

The Polygraphic Record Sets

Preparation.--Following the selection of the sample all record sets were prepared for use in the study. Such preparation consisted of obscuring from the records all

Table 3.2.--Characteristics of Record Sets.

Characteristic of Record Set	Categories of Record Sets							
	Verified				Unverified			
	Truthful		Deceptive		Truthful		Deceptive	
	Crimes Against A Person	Property Crimes	Crimes Against A Person	Property Crimes	Crimes Against A Person	Property Crimes	Crimes Against A Person	Property Crimes
No. CQ Tests*								
Basic Battery only	7	7	10	10	10	6	10	4
Basic Battery plus	7	7	4	4	4	8	4	10
Dual Respiration Recorded								
Yes	4	2	3	5	6	1	2	3
No	10	12	11	9	8	13	12	11
"Yes" Test								
Yes	0	0	7	10	0	1	12	6
No	14	14	7	4	14	13	2	6

* Not counting "yes" Tests.

writing which identified either the subject or examiner, as well as any other notations not pertinent to the numbering of questions asked, the subject's answers, or adjustments to the recordings. Each subject's tests were then arranged in the sequence in which they were given by the examiner, the sequence for all subjects consisting of at least CQ test #1, the "card" test, and CQ test #2. In cases where additional tests had been conducted they were properly placed in the sequence, the only exception being "yes" tests. These tests were eliminated from the study for two reasons. First, the interpretation of the "yes" test is not consistent with the interpretation of other tests, such "response" data not being evaluated in the same manner as that of other tests. Second, the majority of all subjects in the sample who were given "yes" tests were indicated as "deceptive". Hence, it was believed that by excluding the "yes" tests, the evaluators would not, by noting the mere presence of such a test, be able to infer that a given subject was "deceptive" without having to consider response data.

After the masking of extraneous data in the records, all tests for each subject were stacked one on the other, with CQ Test #1 on top of the "card" test, the "card" test on top of CQ #2, etc. The initial portion of the chart paper, preceeding the inflation of the cardio-cuff in each test, was folded and sandwiched between two cardboard retainers which were then securely stapled together. Thus, the

records (tests) of each of the 112 subjects were bound together in the order in which they had been administered; there were 112 record sets consisting of all tests given to each of the subjects, with the exception, as previously noted, of "yes" tests.

Identification.--As will be discussed shortly, it was necessary to distribute the record sets to evaluators over a period of time. For this reason, after the sample was selected all sets were randomly assigned to one of four groups, A, B, C, D, of 28 sets each. The cardboard retainers for each set were then permanently assigned an alphabetic letter corresponding to the group to which the set was assigned, A, B, C, D. Then forty randomly selected sets, five from each of the eight categories originally drawn, were further identified by assignment of the letters QC to the retainers. Finally, adhesive labels were affixed to the retainers on each set; the labels were consecutively numbered 1-28 in group A, 29-56 in group B, etc.

Distribution scheme.--The main reason for division of the sample of record sets into four smaller groups was to facilitate distribution to evaluators. The use of smaller groups also interfered less with the normal workload of the evaluators and enabled several evaluators to be engaged in the study at the same time.

All sets assigned to each group (A,B,C,D) were placed in individual envelopes which in turn were placed into larger envelopes or packets; hence, there were 4 packets each containing 28 record sets, each in individual envelopes. Packets were rotated among evaluators until all had reviewed the records in all four packets. Each evaluator was asked to complete his review of a packet within one week of receiving it.

The distributional scheme mentioned above necessitated insuring reduction, if not elimination of collaboration between evaluators. This was accomplished by the use of adhesive labels identifying record sets and by which the identification number on sets in any one packet were readily altered. After an evaluator at one location completed his review of the sets in a packet, the packet along with the evaluator's "answer sheets" were returned to the researcher. At that time, the record sets in that packet were given new identification numbers.[19] All sets within a packet were consistently ascribed the same array of 28 numbers; the numbers, however, did not correspond to the same record sets. The alteration of identification numbers was done in accordance with a predetermined code sheet by which identification numbers could be matched with the actual subject whose records were used.

---

[19] In two instances when two evaluators were stationed at the same location, identification numbers were not changed between the evaluators' reviews.

After all the sets in a packet were assigned new identification numbers, the packet was then forwarded to another evaluator for review. The second evaluator, because of the altered identification number was uninfluenced by contact with evaluators already familiar with the records.

The Evaluators

Ten polygraphic examiners volunteered to serve as evaluators in the study, all of them, at the start of the study employed by the SPD.[20] As noted earlier, however, the evaluators were stationed at different locations throughout the state; hence the need for the distributional scheme mentioned earlier.

Polygraphic training.--All evaluators had received their initial training in polygraphy at one of two nationally recognized training schools, both schools teaching the application of the Control-Question Technique. There is good reason to believe that the evaluators are representative of all persons trained at these schools for law enforcement agencies, at least in respect to general ability and experience in police or other investigative work.[21]

---

[20]The SPD employs 14 examiners, four of whom were unable to take part in this study. One evaluator retired from the SPD shortly after the start of this study.

[21]Based on personal correspondence between the writer and R. Arther, Director, National Training Center of Lie Detection and L. Marcy, Director, The American Institute of Polygraph Technology and Applied Psychology.

One distinguishing feature concerning the training of the evaluators, however, concerns the internship following their initial training.[22] Such internship consisted of conducting about 200 polygraphic examinations, usually taking between 9 and 15 months, under the personal supervision of qualified examiners within the SPD. The internship, of course, is designed to assure intern examiner's familiarity with the testing procedure used by the SPD examiners and to determine his interest and ability in lie detection. This type of internship is not necessarily a part of the polygraphic training of most examiners employed by law enforcement agencies.[23]

Preparation for study.--Prior to the start of the study a briefing session was held with all evaluators. At that time all were told about the general nature of the study but were not told of the specific hypotheses of interest or the breakdown of record sets involved. Rather, they were told merely that the record sets constituted a representative sample of examinations conducted by the SPD. And, of course, during the session evaluators were given instructions concerning how the record sets were to be distributed and

---

[22]At the start of the study two evaluators were completing the internship.

[23]Such internship is also, in a few states, legally required. See: C. Romig, "The Status of Polygraph Legislation of the Fifty States," Police, 16 (1971), 35-41.

what the nature of their judgments would be; that is, the extent and type of evaluations they would be required to make for each record set.

It is important at this time to mention that not all of the evaluators had been specifically trained in the use of the numerical evaluation-system used for scoring response data, although all were familiar with it. In order to insure that all evaluators understood the system, however, it was reviewed during the briefing session. Moreover, instruction sheets were included in each packet.[24]

Background characteristics.--During the briefing session evaluators completed a background questionnaire indicating their age, years of police investigative experience, years of experience in conducting polygraphic examinations, the approximate number of examinations conducted, and the training school attended. These background data are displayed in Table 3.3.

In Table 3.3 evaluators have been arbitrarily categorized with respect to their years of experience in polygraphic testing; high experience evaluators having three or more years of such experience, low experience evaluators less than three years. The mean age for the former group was 45.4; they had an average of about 5.1 years of experience in polygraphic testing and had conducted an average 1541

---

[24]See Appendix B.

Table 3.3.--Background Characteristics of Evaluators.

Evaluator Experience Level*		Background Data						
		Age	Police Investigative Experience (yrs)	(mos)	Polygraphic Testing Experience (yrs)	(mos)	Approx. no. Polygraphic Examinations Conducted	Training School Attended**
	E_1	38	15	6	1	1	225	A
	E_2	41	7	0	0	9	110	B
Low	E_3	35	13	4	0	8	110	B
	E_4	33	5	0	1	0	227	A
	E_5	41	6	0	2	0	350	A
	E_6	52	14	0	5	5	1450	A
High	E_7	44	8	0	6	2	2500	A
	E_8	42	14	5	3	7	1050	A
	E_9	46	25	0	7	0	1705	A
	E_{10}	43	18	0	3	6	1000	A

* The ordering of evaluators in this table does not correspond to any other table.
**A=National Training Center of Lie Detection, New York, New York.
B=The American Institute of Polygraph Technology and Applied Psychology, Dearborn, Michigan.

polygraphic examinations. On the other hand, low experience evaluators had a mean age of 37.6, an average of 1.1 years of experience in polygraphic testing, and had conducted an average of 204 examinations.

Operational Measures

Evaluators were requested to make several judgments concerning each record set. Such judgments were indicated on two separate answer sheets: an Evaluator Answer Sheet and a Numerical Evaluation Score Sheet. Specimen copies of each of these are displayed in Appendix C.

An Evaluator Answer Sheet was completed by each evaluator for each of the 112 record sets. On this sheet each evaluator indicated his judgment of "truthfulness-deception" indications, "confidence" and "ease of interpretability" of the three basic physiological measures.

Accuracy scores.--The truthfulness-deception judgment was a tripartite one; that is, each evaluator reviewed each record set blind, i.e., without any knowledge of the characteristics of the subject from whom the records were obtained or the nature of the investigation, and decided if it indicated truthfulness (NDI: no deception indicated to relevant questions), deception (DI: deception indicated to relevant questions), or was inconclusive, (INC: response data did not allow for a determination). Since all evaluators were familiar with the standard notational system used for

indicating the various question-types it was unnecessary to identify these in the record sets. Moreover, evaluators were told that their truthfulness-deception judgments were to be based on the complete record set for each subject and that any system of evaluation (visual inspection or numerical evaluation) could be used in forming such a judgment.

The accuracy of truthfulness-deception judgments was of particular interest in the study. Thus, for such judgments made on verified record sets accuracy (correct judgments) was defined as agreement with the known truthfulness or deception of the subject from whom the records had been obtained, using a confession as the criterion measure. It is obvious that such a criterion was unavailable when considering unverified record sets. Hence, judgments made on these sets were defined as correct if the evaluator's judgment agreed with that of the testing examiner. By definition all inconclusive judgments were incorrect.

Since there were eight categories from which record sets were drawn it was possible for each evaluator to make 14 correct judgments within each category. These raw number scores were, however, transformed to percentages; hence, accuracy-scores refer to the percentage of correct judgments made.

Confidence scores.--Each evaluator indicated the degree of confidence in his truthfulness-deception judgment

on a six-point scale ranging from no-confidence (1) to almost-certain (6). The scale was similar to that used by Kubis[25] and Moroney[26] in studies of experimental lie detection.

Confidence scores for each evaluator were defined as the sum of the values, or ratings, indicated on the scale for all record sets within each of the eight categories from which the sets were drawn. Hence, for each evaluator such scores had a theoretical range of 70 points (varying from 14-84) in each category, higher scores indicating greater confidence in the judgments made.

Ease of interpretability scores.--For each record set evaluators rated the "ease of interpretability" of each of three physiological measures: respiration (abdominal respiration only where a dual recording was apparent), GSR, and cardiovascular activity. Such ratings were indicated on a five-point scale for each measure ranging from "very difficult" (1) to "very easy" (5). Again, the scale used was similar to that of Kubis.[27]

---

[25]J. Kubis, Studies in Lie Detection: Computer Feasibility Considerations, Tech. Report 62-2-5 (Arlington, Va.: Armed Services Technical Information Agency, June, 1962), prepared for Air Force Systems Command, Contract No. AF 30 (602)-22700, Project No. SS34, Fordham University, 1962, 146.

[26]W. Moroney, "The Detection of Deception as a Function of PGR Methodology" (unpublished Ph.D. dissertation, St. John's University, 1968, Ann Arbor, Michigan: University Microfilms, 1969, No. 69-7125).

[27]Kubis, Studies in Lie Detection: Computer Feasibility Considerations, op. cit., 146.

There were, in effect, four ease-of-interpretability scores: one for each individual physiological measure with a theoretical range of 56 points (14-70) per category and one for a total ease-of-interpretability considering the three individual scores collectively. The latter score had a theoretical range of 168 points (42-210) in each category. In all cases higher scores indicated greater ease-of-interpretability.

Numerical evaluation.--The numerical evaluation score sheet was completed by evaluators for the forty record sets which had been identified by a QC on the cardboard retainers.[28] On this sheet evaluators were required to assign a number on a 7-point scale ranging from -3 to +3 for each of three physiological measures to indicate the perceived difference between each of four relevant-control question pairings in each of two control question tests. A score of -3 to one of the control-relevant question pairs for each measure indicated a dramatically greater response to the relevant question in that pair; a score of +3 indicated a dramatically greater response to the control question.

To assure that evaluators consistently paired (and scored) the same relevant-control questions, all such pairs were pre-determined and indicated on the numerical evaluation

---

[28]The letters QC refer to "quality control" which is sometimes used synonymously with numerical evaluation although, they are not, in fact, identical concepts. See: R. Brisentine, "Quality Control," Polygraph, 2 (1973), 278-286.

sheet. Moreover, as discussed earlier, evaluators were required to score only abdominal respiration in those instances where a dual respiratory recording was evident and only CQ test #1 and CQ test #2 in those instances where additional tests were included in a record set.

There were eight basic scores generated for each evaluator for each record set numerically evaluated: a score for each of three measures and a total score (the algebraic sum of the individual scores for the three measures), for each of two tests. However, for purposes of the study such scores were combined in the following manner: a score for each of the three measures was obtained by algebraically summing the scores for each measure for the two tests; a combined score for all measures was derived by algebraically summing the total scores for both tests. Hence, there were four scores obtained for each record set numerically evaluated by each evaluator: a score for each of three measures (physiological components) each with a theoretical range from +24 to -24 and a combined score for the record set with a theoretical range from +72 to -72.

Evaluator experience.--Evaluators were categorized as high-experience, more than 3 years of experience in conducting polygraphic examinations, and low experience, less than 3 years. Although such categorization was arbitrary, it will be noted on inspection of Table 3.3, page 107, that the criterion naturally sorted the evaluators into two equal groups.

Hypotheses

Hypothesis-testing procedures were carried out for a series of research hypotheses developed with respect to accuracy scores, confidence scores and ratings, and total ease-of-interpretability scores and ratings. These hypotheses are presented below along with a summary of their rationale.

Accuracy Scores

> Hypothesis I: High-experience evaluators will attain higher accuracy scores than low-experience evaluators.

Rationale.--Horvath and Reid and Hunter and Ash have reported that experienced evaluators are more accurate (and consistent) in their judgments of polygraphic records than less experienced evaluators; Hypothesis 1 is consistent with these investigators' findings.

> Hypothesis II: Accuracy-scores on record sets drawn from verified investigations will be higher than those on sets drawn from unverified investigations.

Rationale.--Verified investigations are those where the testing examiner correctly identified the guilty person. It is argued that such identification depended upon an appropriate pre-test interview, stimulation strategies, etc., which in turn led to clearly recognizable physiological responses. Thus, Hypothesis II is based on the assumption that record sets drawn from verified investigations are more dependable than those drawn from unverified investigations.

Hypothesis III: Accuracy-scores on record sets of truthful subjects will be higher than those on sets of deceptive subjects.

Rationale.--Hypothesis III is consistent with the findings of Horvath and Reid, and the claims of many field-examiners, that errors are made more often on deceptive than truthful subjects; that is, "false negatives" occur more often than "false positives".

Hypothesis IV: Accuracy-scores on record sets drawn from investigations concerning crimes against a person will be higher than those on sets concerning property crimes.

Rationale.--Hypothesis IV is based upon the assumption that when testing subjects involved in crimes against a person, an examiner has access to more detailed information concerning the offense than is typically available in property-crime investigations. Such detailed information leads to more appropriate question-formulation and thus more clearly recognized physiological responses.

Confidence Scores

Hypothesis V: High-experience evaluators will attain higher confidence scores than low-experience evaluators.

Rationale.--It is not known if experienced evaluators have greater confidence in their judgments than do inexperienced. It is reasonable to suspect that they do, particularly in view of Horvath and Reid's suggestion that experience enables an evaluator to apply consistently the "fine points" of the theory of control-question testing when making judgments.

Hypothesis VI: Confidence-scores will be higher for judgments made on record sets drawn from verified investigations than for those made on sets from unverified investigations.

Rationale.--Hypothesis VI is based on the assumption that physiological data in record sets drawn from verified investigations are more dependable than in those from unverified investigations. In other words, confidence will increase when more clearly recognizable physiological responses are apparent.

Hypothesis VII: Confidence-scores will be higher for judgments made on record sets of truthful subjects than those of deceptive subjects.

Rationale.--Field examiners maintain that truthful subjects are easier to detect than deceptive subjects, presumably because of clearer response patterns. Hence, confidence scores will be greater in such judgments.

Hypothesis VIII: Confidence-scores will be higher for judgments made on record sets drawn from investigations concerning crimes against a person than those concerning property crimes.

Rationale.--Assuming that response-patterns are more clearly recognizable when considering record sets of subjects involved in crimes against a person, confidence scores will be greater for judgments of such records.

Hypothesis IX: Confidence-ratings will be higher for correct than for incorrect judgments.

Rationale.--Kubis reported that evaluators of experimentally derived polygraphic records had greater confidence in correct than in incorrect judgments. Hypothesis IX is consistent with Kubis's findings.

Ease of Interpretability Scores

Hypothesis X: High-experience evaluators will have higher total ease-of-interpretability scores than will low experience evaluators.

Rationale.--If, as Horvath and Reid suggest, experience enables an evaluator to apply consistently the fine points of the theory of control question testing, it is reasonable to suspect that experienced evaluators will report polygraphic records easier to interpret than less experienced evaluators.

Hypothesis XI: Total ease-of-interpretability scores will be higher in judgments of record sets drawn from verified investigations than those made on sets drawn from unverified investigations.

Rationale.--Assuming that verified records consist of more clearly recognizable physiological responses, they will be judged easier to interpret than unverified records.

Hypothesis XII: Total ease-of-interpretability scores will be higher in judgments of record sets of truthful subjects than those of deceptive subjects.

Rationale.--Field examiners maintain that the polygraphic records of truthful subjects are easier to interpret than those of deceptive subjects. Hypothesis XII is consistent with this claim.

Hypothesis XIII: Total ease-of-interpretability scores will be higher in judgments of record sets drawn from crimes against a person than those of sets drawn from property crimes.

Rationale.--Hypothesis XIII is consistent with the assumption that clearer response-patterns are evident in those records drawn from subjects involved in crimes against a person than in property crimes.

> Hypothesis XIV: Total ease-of-interpretability ratings will be higher for correct than for incorrect judgments.

Rationale.--Hypothesis XIV is consistent with Kubis's findings that records on which correct judgments were made were easier to interpret than those judged incorrectly.

Design and Analysis

The design used for hypotheses testing, except with respect to hypotheses #IX and #XIV, was a 2 . 2 x 2 x 2 Split-plot (repeated measures) described by Kirk as type SPF P .qru.[29] A dummy data matrix defined in terms of the study is shown in Figure 3.2.

Using the design indicated in Figure 3.2 a four-way Analysis of Variance (ANOVA), repeated measures, was carried out to simultaneously test appropriate (null) hypotheses with respect to the (research) hypotheses developed for each of three dependent measures generated, accuracy scores, confidence scores, and total ease-of-interpretability scores. The four factors were: Verification (verified and unverified); Truthfulness (truthful and deceptive); Crime-type (crimes

---

[29] R. Kirk, Experimental Design: Procedures for the Behavioral Sciences (Belmont, Calif.: Brooks/Cole, 1968), 308.

Evaluator Experience Level		Categories of Record Sets							
		b_1	b_1	b_1	b_1	b_2	b_2	b_2	b_2
		c_1	c_1	c_2	c_2	c_1	c_1	c_2	c_2
		d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
	e_1								
	e_2								
a_1	e_3								
	e_4								
	e_5								
	e_6								
	e_7								
a_2	e_8								
	e_9								
	e_{10}								

A = Experience (a_1=low, a_2=high)
B = Verification (b_1=verified, b_2=unverified)
C = Truthfulness (c_1=truthful, c_2=deceptive)
D = Crime type (d_1=person, d_2=property)
e = evaluators:A

Figure 3.2--Dummy Data Matrix: 2.2x2x2 Split-plot.

against a person and property crimes) and Experience of evaluators (high and low), the first three treated as repeated measures.

The testing of appropriate null hypotheses for hypotheses #IX and XIV was carried out with two-way ANOVA, repeated measures, in a 2.2 Split-plot design as shown in Figure 3.3.

The two factors were: Evaluator-experience (high and low) and Judgments (correct and incorrect), treated as repeated measures. Dependent variables treated separately using this design were mean confidence ratings and mean total ease-of-interpretability ratings for correct and incorrect judgments.

In all instances the .05 level of significance was established as the decision rule regarding the testing of hypotheses. That is, null hypotheses were rejected when the probability of a Type I error was equal to or less than .05.

Evaluator Experience	Level	Judgments: Correct	Judgments: Incorrect
	e_1		
	e_2		
Low	e_3		
	e_4		
	e_5		
	e_6		
High	e_7		
	e_8		
	e_9		
	e_{10}		

Figure 3.3--Dummy Data Matrix: 2.2 Split-plot.

To determine the reliability of the numerical scoring system the Pearson product-moment correlation coefficient (r) was used. Such correlations were calculated for the set of scores between all possible pairs of evaluators for each of the four numerical scores generated, respiration, GSR, cardio, and combined scores, for the record sets.

Analysis of data other than that explained above is more appropriately described in the next chapter.

Chapter IV

RESULTS

Accuracy of Judgments

Overall, the ten evaluators made 1120 truth/deception judgments; of these, 707, or 63.1 percent, were correct ($p < .001$).[1] The discard of the fifteen "inconclusive" judgments made by the evaluators was not sufficient to substantially alter the grouped results.

Accuracy-scores for individual evaluators in each of the eight categories of record sets are displayed in Table 4.1; as indicated, the total accuracy-scores for the low-experience evaluators ranged from 61.6 to 64.3 percent, for the high-experience evaluators from 53.6 to 69.6 percent.[2] The evaluator with the lowest total accuracy score also made the greatest number of "inconclusive" judgments, which, as pointed out in Chapter III, were scored as errors; were these "inconclusives" eliminated, this evaluator's score would be consistent with other evaluators' scores.[3]

---

[1]Using the binominal approximation to the normal distribution and treating the data as though there were two legitimate outcomes, correct and incorrect.

[2]The raw numbers on which these percentages are based are displayed in Appendix D, Table D.1.

[3]This evaluator reported eight "inconclusive" judgments, one more than all such judgments made by all other evaluators.

TABLE 4.1.--Accuracy Scores of Individual Evaluators

Evaluators	Categories of Record Sets								Total Percent Correct Judgments
	Verified				Unverified				
	Truthful		Deceptive		Truthful		Deceptive		
	Crime Against Person	Property Crime	Crime Against Person	Property Crime	Crime Against Person	Property Crime	Crime Against Person	Property Crime	
Low Experience	%	%	%	%	%	%	%	%	%
E_1	64.3	64.3	78.6	57.1	64.3	71.4	71.4	42.9	64.3
E_2	35.7	35.7	92.9	100.0	21.4	35.7	92.9	78.6	61.6
E_3	50.0	35.7	78.6	85.7	21.4	64.3	92.9	78.6	63.4
E_4	35.7	42.9	92.9	78.6	14.3	71.4	71.4	85.7	61.6
E_5	50.0	64.3	78.6	78.6	35.7	64.3	85.7	42.9	62.5
Sub-Total	47.1	48.6	84.3	80.0	31.4	61.4	82.9	65.7	62.7
High Experience									
E_6	42.9	57.1	92.9	85.7	57.1	71.4	78.6	71.4	69.6
E_7	57.1	35.7	92.9	71.4	28.6	57.1	85.7	85.7	64.3
E_8	64.3	64.3	71.4	71.4	35.7	78.6	78.6	50.0	64.3
E_9	57.1	71.4	28.6	50.0	42.9	71.4	64.3	42.9	53.6
E_{10}	42.9	50.0	78.6	78.6	57.1	78.6	78.6	64.3	66.1
Sub-Total	52.9	55.7	72.9	71.4	44.3	71.4	77.1	62.9	63.6
TOTAL	50.0	52.1	78.6	75.7	37.9	66.4	80.0	64.3	63.1

Hypotheses

Main effects.--A four-way analysis of variance (ANOVA), repeated measures, was conducted on the individual accuracy-scores shown in Table 4.1. The four factors were: Verification (verified and unverified); Truthfulness (truthful and deceptive); Crime Type (crime against a person and property crime), all treated as repeated measures; and Experience of evaluators (high and low).

The Hypotheses formulated with respect to accuracy-scores are presented below, along with the results of the ANOVA.[4]

> Hypothesis I: High-experience evaluators will attain higher accuracy-scores than low-experience evaluators.

Although overall the high-experience evaluators did attain higher accuracy-scores than did the low-experience group (63.6 and 62.7 percent correct, respectively), Hypothesis I is not supported by the results of the ANOVA. The main effect pertaining to differences between groups (experience levels) of evaluators with respect to accuracy-scores was not significant [$F(1,8)=.11$, $p > .10$]. There were no significant interaction-effects involving the experience-groupings of evaluators.

---

[4]All ANOVA tables not in the text are displayed in Appendix E; Table E.1 details the ANOVA results for accuracy scores.

Hypothesis II: Accuracy-scores on record sets drawn from verified investigations will be higher than those on sets drawn from unverified investigations.

The accuracy-scores of both groups of evaluators combined indicate that 64.1 percent of the judgments made on verified record sets were correct, as opposed to 62.1 percent on the unverified sets. These data are shown in Table 4.2. The difference in accuracy between verified and unverified record sets was in the predicted direction but was not significant [$F(1,8)=1.42$, $p > .10$]. Apparent, however, was a significant interaction effect involving the verification categories, an effect to be discussed later in this paper.

TABLE 4.2.--Accuracy on Record Sets in Verified and Unverified Categories.

Evaluator Experience Level	Category	
	Verified	Unverified
Low	65.0%	60.4%
High	63.2%	63.9%
Combined	64.1%	62.1%

Hypothesis III: Accuracy-scores on record sets of truthful subjects will be higher than those on sets of deceptive subjects.

Hypothesis III is not supported. As shown in Table 4.3, the evaluators were correct in 51.6 percent of their

judgments on "truthful" record sets, and 74.6 percent on "deceptive". This difference was significant [F (1,8)=10.70, p< .01], and contrary to the predicted direction. Two significant interaction effects complicating the meaning of this result will be subsequently discussed.

TABLE 4.3.--Accuracy on Record Sets in Truthful and Deceptive Categories.

Evaluator Experience Level	Category	
	Truthful	Deceptive
Low	47.1%	78.2%
High	56.1%	71.1%
Combined	51.6%	74.6%

Hypothesis IV: Accuracy-scores on record sets drawn from investigations concerning crimes against a person will be higher than those on sets concerning property crimes.

Classification of record sets by type of crime (Table 4.4) indicates that the evaluators were correct in 61.6 percent of their judgments on record sets concerning "crimes against a person" and 64.6 percent on those concerning "property crimes". This result contradicted the predicted direction but not to a statistically significant extent [F (1,8)=1.54, p> .10]. Two interaction effects involving the classification of record sets by type of crime are discussed below.

TABLE 4.4.--Accuracy on Record Sets Classified by Type of Crime.

Evaluator Experience Level	Crime Classification	
	Crime Against a Person	Property Crime
Low	61.4%	63.9%
High	61.8%	65.4%
Combined	61.6%	64.6%

Interaction effects.--The ANOVA conducted on accuracy scores revealed two significant interaction effects, a Truthfulness x Crime type [$F(1,8)=55.83$, $p < .001$] and a Verification x Truthfulness x Crime type interaction [$F(1,8)=20.87$, $p < .002$]. A discussion of these effects, considering first the two-way interaction, follows.

Figure 4.1 displays the means for the Truthfulness x Crime type interaction. As shown, these means plot ordinally, the record sets of deceptive subjects being judged correctly more often than those of truthful subjects, irrespective of crime. The higher-order interaction, however, complicates the meaning of the data regarding the two-way interaction.

Figure 4.2 displays the mean-accuracy scores for the Verification x Truthfulness x Crime type interaction. Inspection of this figure shows that record sets in the crime-against-

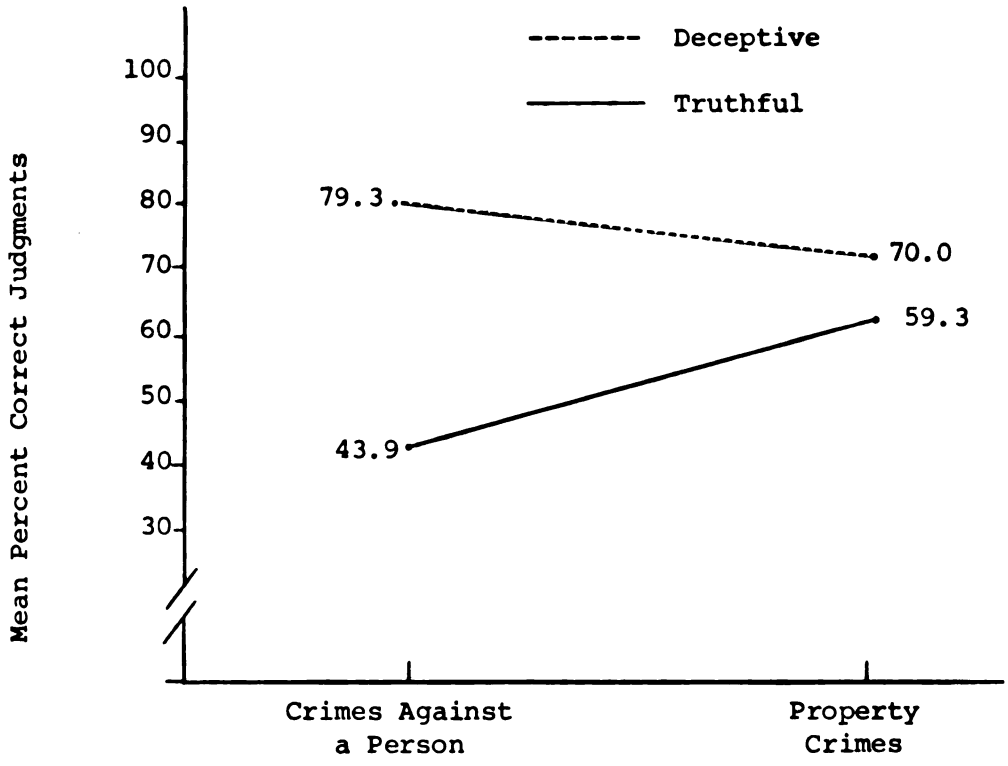

Figure 4.1.--Mean percent correct judgments on record sets in the truthful and deceptive categories in the two crime classifications.

a-person classification were correctly judged more often than all others if they were also in the deceptive category, less often if they were in the truthful category, regardless of the verification. It can also be seen that there was an ordinal effect considering only the crime-against-a-person classification, record sets in the deceptive category being correctly judged more often than those in the truthful category whether verified or unverified. A disordinal relationship obtained considering only the property-crime classification; record sets in the deceptive category of this classification were correctly judged more often than those in the truthful category only in the verified condition.

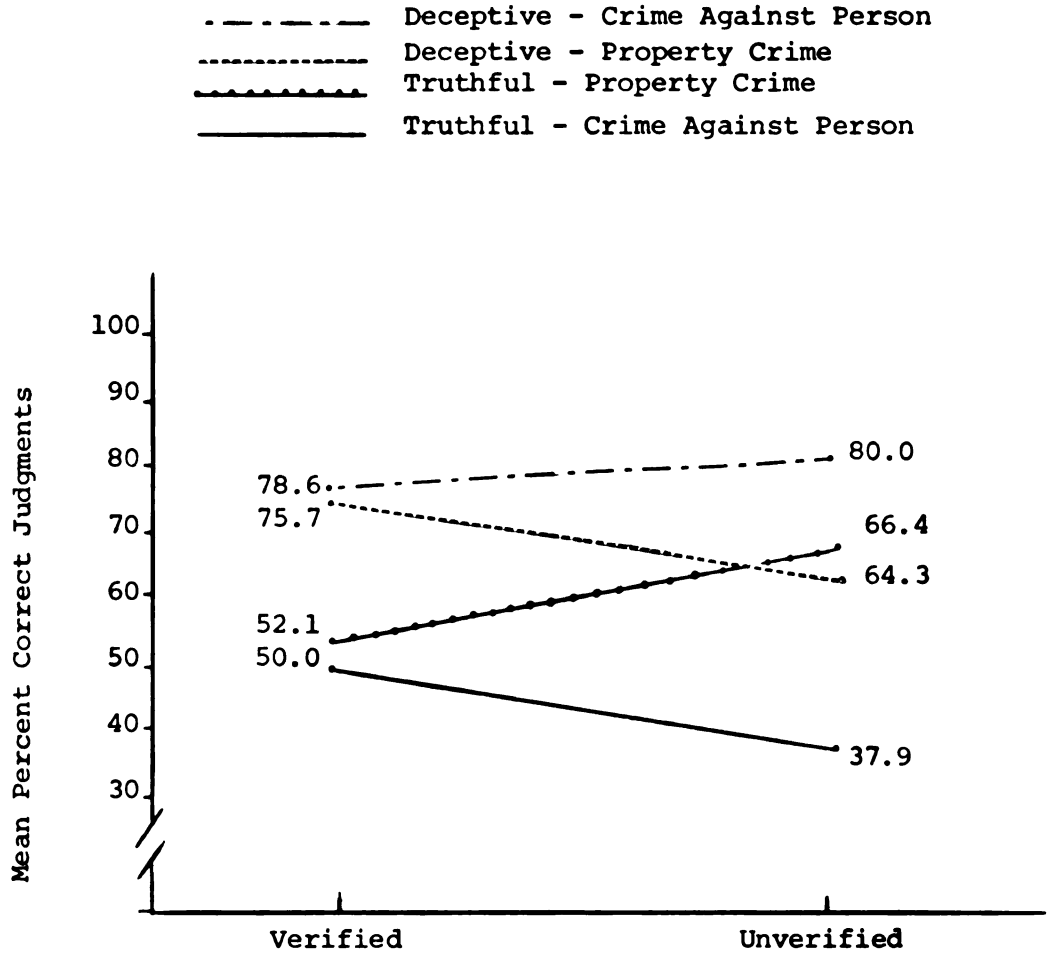

Figure 4.2--Mean percent correct judgments on deceptive and truthful crime against a person and property crime record sets for the verified and unverified conditions.

Collective Accuracy

While no predictions were made with respect to the accuracy of collective judgments of evaluators such accuracy will be briefly discussed here. There were 104 record sets on which six or more evaluators made definitive judgments of truthfulness or deception. When six evaluators agreed, collective judgments were correct in three of thirteen

(23.1 percent) such occurrences; when all ten evaluators agreed, eighteen of twenty-one (85.7 percent). When agreement between six or more evaluators obtained, sixty-seven of the 104 such agreements were correct (64.4 percent). These data, along with the accuracy of the intermediate levels of evaluator-agreement are shown in Table 4.5, which also shows that there was a positive relationship between collective accuracy and the number of evaluators in agreement in their truth/deception judgments.

TABLE 4.5.--Accuracy of Collective Judgments of Evaluators.

	Judgments			
Number of Evaluators Agreeing	Correct		Incorrect	
	No.	(%)	No.	(%)
6	3	(23.1)	10	(76.9)
7	10	(55.6)	8	(44.4)
8	17	(68.0)	8	(32.0)
9	19	(70.4)	8	(29.6)
10	18	(85.7)	3	(14.3)
TOTAL	67	(64.4)	37	(35.6)

Effect of Additional Physiological Data

As explained in Chapter III, the record sets used in this study were not uniform in nature, some containing Control-Question tests beyond the basic battery (CQ test #1, "card" test, CQ test #2) and some recorded by a polygraphic

instrument with dual respiration-components. Although no predictions were made concerning the effect which these variables would have on the accuracy of judgments, it is of some interest to examine this effect. The percentage of each evaluator's correct judgments for two conditions for each variable was calculated. Using these percentages as a dependent-variable, t-tests for correlated means ($t_{dep.}$) were conducted to determine if there was a significant difference in accuracy between the two conditions for each variable. It should be noted, however, that the variables themselves were not necessarily independent.

Table 4.6 compares the mean percentage of correct judgments on record sets containing a dual respiration tracing to those sets of only a single such tracing. While the table shows percentages for both groups of evaluators, the groups were not treated as a factor in the analysis. As indicated in Table 4.6 correct judgments were made an average of 67.7 percent when a dual tracing was apparent, 61.7 percent when a single tracing was used. Although the accuracy was higher for the first condition, the difference was not significant ($t_{dep.}=1.84$, $p< .10$); a two-tailed test was used since no predictions were made concerning this variable.

Table 4.7 displays the mean accuracy of judgments when record sets are dichotomized, with respect to the number of CQ tests, containing only the basic battery of tests

and the basic battery plus additional tests. The mean accuracy on record sets in the former category was 71.1 percent, in the latter, 52.5 percent. This difference was significant ($t_{dep.}$=8.21, p< .001), when a two-tailed test was used.

TABLE 4.6.--Accuracy of Judgments Based on Number of Respiration Components Recorded.

Evaluator Experience Level	Dual Respiration Recorded	
	Yes	No
Low	64.6%	62.1%
High	70.8%	61.4%
Combined	67.7%	61.7%

t dep. = 1.84, df= 9, p.<.10

TABLE 4.7.--Accuracy of Judgments Based on Number of Control Question Tests in Record Sets.

Evaluator Experience Level	Number Control Question Tests	
	Basic Battery Only	Basic Battery +
Low	68.4%	55.0%
High	73.8%	50.0%
Combined	71.1%	52.5%

t dep. = 8.21, df = 9, p.<.001.

Reliability of Judgments

The extent of agreement of all evaluators on all record sets irrespective of the correctness of judgments, is apparent from inspection of the data presented in Table 4.5, page 129. It is of interest to examine these data and evaluator-reliability in greater detail.

Of the 112 record sets, 104 were agreed upon by six or more evaluators, as indicative of either truthfulness or deception. In the eight instances where such agreement was not apparent, five were even splits (five evaluators making judgments of truthfulness and five of deception). In the remaining three instances, inconclusive judgments were rendered by one or more evaluators, precluding majority agreement, because of the distribution of definitive judgments.

To determine the extent of inter-evaluator agreement, irrespective of accuracy, the percentage of agreements in judgments between all possible pairs of evaluators were calculated; since there were ten evaluators, forty-five pairings were possible. These percentages, displayed in Table 4.8, ranged from 53 to 90 percent, with a mean of 69 percent. In other words, two evaluators agreed on an average of 69 percent of the time that any particular record set indicated truthfulness or deception, or was inconclusive.

Further analysis of the reliability of evaluator-judgments was made by calculating Hoyt's intra-class

TABLE 4.8.--Percentage of Agreements in Paired Judgments of Evaluators.

	Low Experience				High Experience					
Evaluators	1	2	3	4	5	6	7	8	9	10
1	-	60	70	71	65	63	61	68	66	70
2		-	75	65	82	70	74	53	71	68
3			-	81	77	74	72	66	72	74
4				-	67	73	64	66	69	67
5					-	71	79	62	74	75
6						-	68	66	71	71
7							-	54	66	67
8								-	60	60
9									-	90

(reliability) correlation-coefficient for ratings, as described by Ebel.[5] Such correlations were calculated separately for judgments made on verified and unverified record sets by converting all evaluators' judgments to numerical values (1=truthful, 2=inconclusive, 3=deceptive) and conducting a two-way analysis of variance on these values; the two factors were Records (N=56), and Evaluators (N=10). The resulting mean squares were then used to determine reliability-coefficients.[6]

The reliability coefficients for both verified and unverified record sets were quite similar, .89 and .85, respectively, indicating that there was substantial reliability for the ratings (judgments) of all evaluators on record sets in both categories. Said in another way, the variability between the ten evaluators with respect to their judgments of truthfulness/deception indications in the record sets was relatively low.

Confidence in Judgments

Confidence scores were the sum of the values, or ratings, indicated by evaluators on a six-point scale for each record set in each of the eight categories from which

---

[5] R. Ebel, "Estimation of the Reliability of Ratings," in W. Mehrens and R. Ebel (Eds.), Principles of Educational and Psychological Measurement (Chicago: Rand McNally, 1967), 116-131.

[6] In terms of analysis of variance, for this situation, the coefficient was the ratio of the mean square for records minus that for error to the mean square for records.

the sets were drawn. Such scores had a theoretical range of 70 points (14-84) per category, higher scores indicating greater confidence.[7]

Hypotheses

Main effects.--A four-way ANOVA, repeated measures, was conducted to test the main-effect hypotheses formulated with respect to evaluators' confidence-scores. These hypotheses, along with the results of the ANOVA, are presented below.[8]

> Hypothesis V: High-experience evaluators will attain higher confidence scores than low-experience evaluators.

While the high-experience evaluators did report greater confidence in their judgments than the low-experience group, with mean confidence scores of 56.5 and 51.7, respectively, this difference was not significant [$F(1,8)=1.77$, $p > .10$]. Thus, Hypothesis V is not supported. There were no significant interaction-effects associated with experience-levels of evaluators pertaining to confidence scores.

> Hypothesis VI: Confidence-scores will be higher for judgments made on record sets drawn from verified investigations than for those made on sets from unverified investigations.

As indicated in Table 4.9 the mean confidence-scores for all evaluators on record sets in the verified category

---

[7]Confidence scores for individual evaluators are displayed in Appendix D, Table D.2.

[8]The ANOVA table for confidence scores is displayed in Appendix E, Table E.2.

was 54.5, in the unverified category, 53.7. This difference, although in the predicted direction, was not significant [F (1,8)=.53, p> .10]. However, a significant interaction effect involving the Verification factor did emerge from the analysis; this effect will be discussed shortly.

TABLE 4.9.--Mean Confidence Scores on Verified and Unverified Record Sets.

Evaluator Experience Level	Category	
	Verified	Unverified
Low	52.9	50.5
High	56.2	56.9
Combined	54.5	53.7

Hypothesis VII: Confidence-scores will be higher for judgments made on record sets of truthful subjects than those of deceptive subjects.

Table 4.10 displays the mean confidence-scores for both groups of evaluators on record sets in the truthful

TABLE 4.10.--Mean Confidence Scores on Record Sets Classified as Truthful and Deceptive.

Evaluator Experience Level	Category	
	Truthful	Deceptive
Low	49.3	54.1
High	55.0	58.1
Combined	52.1	56.1

and deceptive categories; as shown, for all evaluators the mean in the truthful category was 52.1, in the deceptive category, 56.1. This difference was significant [$F(1,8) = 64.17$, $p < .001$] but opposite the predicted direction, and its meaning is complicated by an interaction effect.

> Hypothesis VIII: Confidence-scores will be higher for judgments made on record sets drawn from investigations concerning crimes against a person than those concerning property crimes.

As predicted, confidence-scores were higher on judgments of record sets in the crime-against-a-person category than in the property-crime category, the mean scores being 54.2 and 54.0, respectively. These data are shown in Table 4.11. However, Hypothesis VIII is not supported by the results of the ANOVA since the difference between the confidence-scores pertaining to crime classification was not significant [$F(1,8) = .08$, $p > .10$]. There were no significant interaction-effects with respect to confidence-scores involving crime classification.

TABLE 4.11.--Mean Confidence Scores on Record Sets Classified by Type of Crime.

Evaluator Experience Level	Crime Classification	
	Crime Against A Person	Property Crime
Low	52.3	51.1
High	56.2	56.9
Combined	54.2	54.0

Interaction effects.--A significant Verification x Truthfulness interaction-effect was apparent in the results of the ANOVA conducted on confidence-scores [F (1,8)=6.23, $p < .03$]. The nature of this interaction can be discerned from inspection of Figure 4.3 which displays the mean confidence-scores for record sets in the truthful and deceptive categories for the verified and unverified conditions.

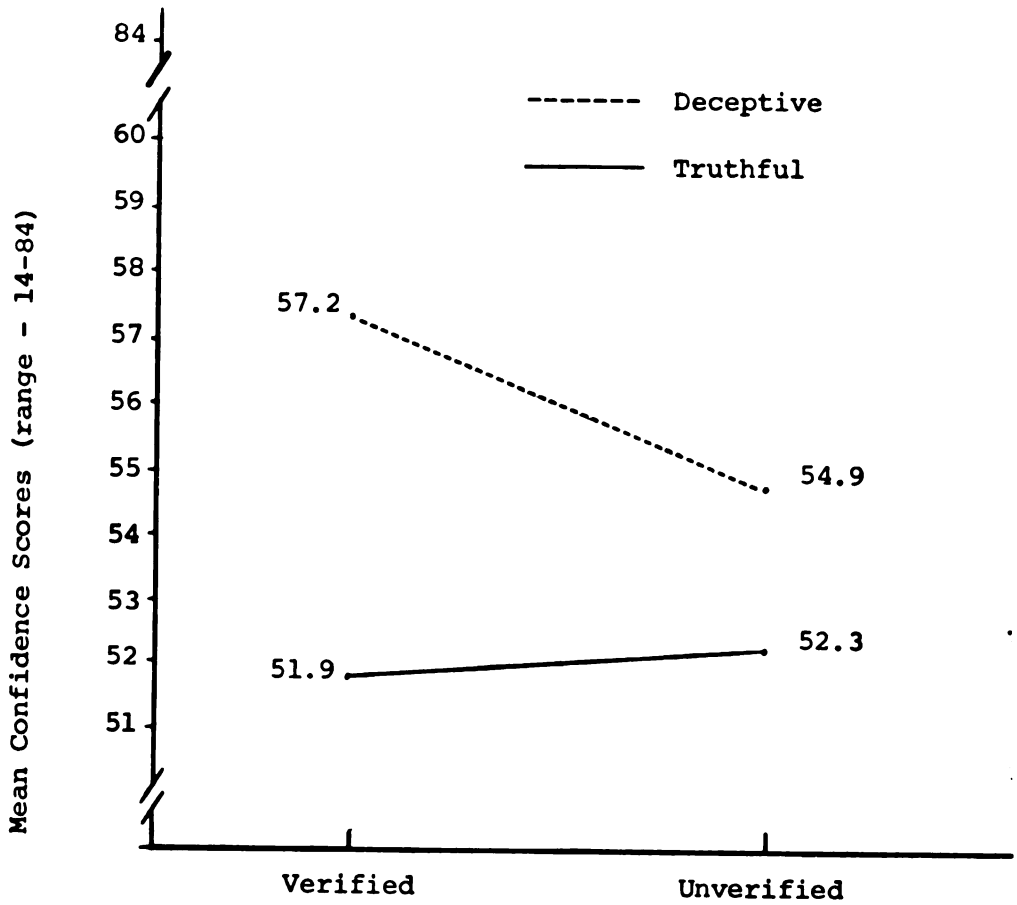

Figure 4.3.--Mean confidence scores on record sets in the deceptive and truthful categories for the verified and unverified conditions.

As shown in Figure 4.3, the interaction mentioned above was ordinal in nature; that is, mean confidence-scores were greater for record sets in the deceptive than the truthful category across the two levels of verification. Moreover, it is apparent that mean confidence-scores decreased from verified to unverified for sets in the deceptive category while they increased slightly for sets in the truthful category.

Confidence Ratings and Accuracy of Judgments

Table 4.12 displays the mean confidence-ratings for both groups of evaluators for correct and incorrect judgments. As shown, the high-experience evaluators had a mean confidence-rating of 4.2 on correct judgments, 3.8 on incorrect; the low-experience evaluators mean ratings of 3.8 and 3.5 for correct and incorrect judgments, respectively.

TABLE 4.12.--Mean Confidence Ratings of Evaluators' Judgments.

Evaluator Experience Level	Judgments	
	Correct	Incorrect
Low	3.8	3.5
High	4.2	3.8
Combined	4.0	3.6

Using each evaluator's mean confidence-rating on correct and incorrect judgments as the dependent variable, a two-way ANOVA, repeated measures, was carried out. The two factors were Judgments (correct and incorrect), treated as repeated measures, and Evaluator-experience (high and low). No prediction was made concerning the main effect for experience levels; Hypothesis IX, however, concerning the main effect for judgments, is presented below, along with the results of the ANOVA.

Hypothesis IX: Confidence-ratings will be higher for correct than for incorrect judgments.

As indicated in Table 4.12, the mean confidence-rating for all evaluators on correct judgments was 4.0; on incorrect, 3.6. As can be seen from inspection of Table 4.13, which details the results of the ANOVA, this difference was significant [$F(1,8)=21.55$, $p< .002$]; Hypothesis IX is supported. The main effect for experience, as also shown in Table 4.13 was not significant, nor was there a significant Experience x Judgment interaction.

TABLE 4.13.--Analysis of Variance Table for Mean Confidence Ratings on Correct and Incorrect Judgments.

Source	SS	df	MS	F	p<
A (A=Experience)	.58	1	.58	1.82	.25
E (E=Evaluators):A	2.54	8	.32		
J (J=Judgments)	.72	1	.72	21.55	.002
A X J	.00	1	.00	0.00	-
J X E:A	.27	8	.03		
TOTAL	4.11	19			

Ease-Of-Interpretability Of Record Sets

For all record sets in each of the eight categories, evaluators rated the ease-of-interpretability of respiration, GSR, and cardiovascular activity. Ease-of-interpretability scores for each component had a theoretical range of 56 points (14-70) per category. A total ease-of-interpretability score for each record set was derived by summing the ratings for individual components; this score had a theoretical range of 168 points (42-210) in each of the eight categories. In all cases, higher scores indicated greater ease-of-interpretability.[9]

Hypotheses

Main effects.--A four-way ANOVA, repeated measures, was conducted on evaluators' total ease-of-interpretability scores.[10] The hypotheses formulated with respect to these scores and the results of the ANOVA are discussed below.

> Hypothesis X: High-experience evaluators will have higher total ease-of-interpretability scores than will low-experience evaluators.

The high-experience evaluators had a mean total ease-of-interpretability score of 116.2, the low-experience group, 106.6. Although this result was in the predicted direction, it was not significant [$F(1,8)=1.03$, $p > .10$]; Hypothesis X

---

[9] Total ease-of-interpretability scores for individual evaluators are displayed in Appendix D, Table D.3.

[10] The ANOVA Table for total ease-of-interpretability scores is displayed in Appendix E, Table E.3.

is not supported. There were no significant interaction-effects associated with the Experience factor regarding total ease-of-interpretability scores.

> Hypothesis XI: Total ease-of-interpretability scores will be higher in judgments of record sets drawn from verified investigations than those made on sets drawn from unverified investigations.

The mean total "ease-of-interpretability" score for all evaluators on record sets in the verified category was 112.9, in the unverified category, 109.9. This difference was significant [$F(1,8)=7.65$, $p< .02$]; therefore, Hypothesis XI is supported. However, because of a significant interaction-effect involving the Verification factor, the meaning of this main effect will be discussed later.

> Hypothesis XII: Total ease-of-interpretability scores will be higher in judgments of record sets of truthful subjects than those of deceptive subjects.

Total "ease-of-interpretability" scores for record sets in the deceptive category had a mean of 115.9; in the truthful, 106.9. This result was significant [$F(1,8)=37.99$, $p< .001$], but opposite the predicted direction; Hypothesis XII is not supported. A significant interaction-effect involving the Truthfulness factor will be discussed shortly.

> Hypothesis XIII: Total ease-of-interpretability scores will be higher in judgments of record sets drawn from crimes against a person than those of sets drawn from property crimes.

Classification of record sets by type of crime shows that the mean total "ease" score for sets in the crime-against-a-person category was 113.7, in the property-crime category, 109.0. This difference was significant [$F(1,8) = 8.22$, $p < .02$], and therefore Hypothesis XIII is supported by the results of the ANOVA. There were no significant interaction-effects associated with classification of record sets by type of crime.

Interaction effects.--A significant Verification x Truthfulness interaction-effect was apparent from the results of the ANOVA conducted on total "ease" scores [$F(1,8) = 9.13$, $p < .02$]. The ordinal nature of this interaction can be seen in Figure 4.4; mean total ease-of-interpretability scores were higher for record sets in the deceptive category than in the truthful category across both levels of the Verification factor. It is also apparent that such scores increased for record sets in the truthful category from the verified to the unverified condition, while they decreased for sets in the deceptive category.

Total Ease-of-Interpretability Ratings and Accuracy

Table 4.14 displays the mean total ease-of-interpretability ratings for both groups of evaluators on correct and incorrect judgments. On correct judgments the mean rating for the low-experience group was 7.8, for the high

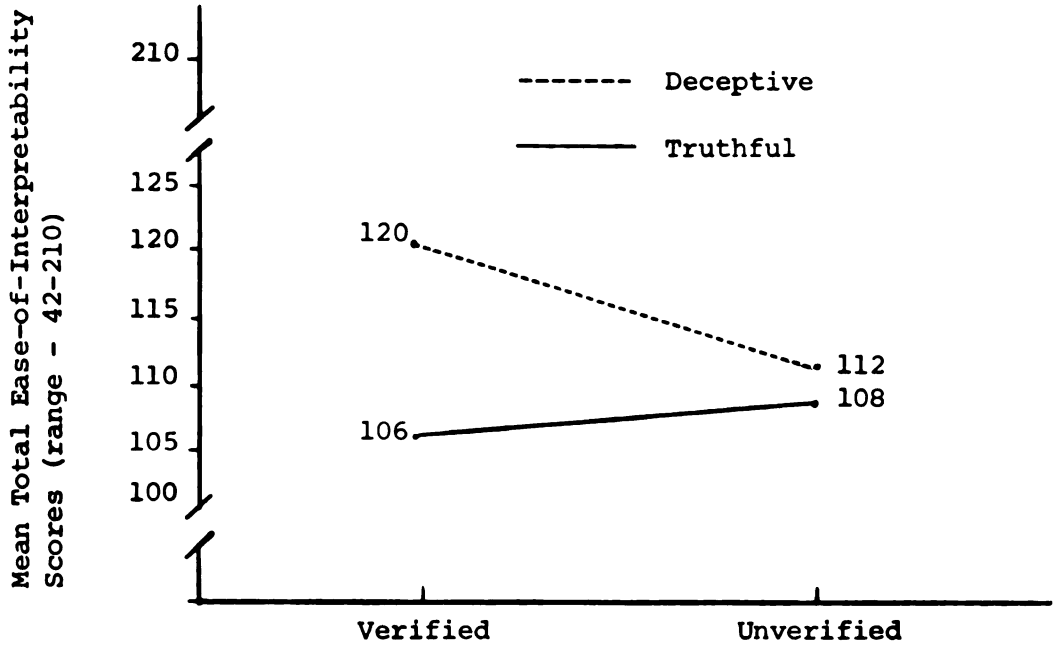

Figure 4.4.--Mean total ease-of-interpretability scores for record sets in the truthful and deceptive categories for the verified and unverified conditions.

experience group 8.5; for each group the mean ratings were higher on correct than incorrect judgments.

TABLE 4.14.--Mean Total Ease-of-Interpretability Ratings of Evaluators' Judgments.

Evaluator Experience Level	Judgments	
	Correct	Incorrect
Low	7.8	7.3
High	8.5	7.9
Combined	8.1	7.6

Using each evaluator's mean total ease-of-interpretability rating on correct and incorrect judgments as the dependent variable, a two-way ANOVA, repeated measures, was carried out. The two factors were Judgments (correct and incorrect) treated as repeated measures, and Evaluator-experience (high and low). No prediction was made concerning the main effect for the Experience factor; the hypothesis pertaining to the main effect for the Judgment factor is discussed below.

> Hypothesis XIV: Total ease-of-interpretability ratings will be higher for correct than for incorrect judgments.

As shown in Table 4.14 the mean total ease-of-interpretability rating for all evaluators on correct judgments was 8.1; on incorrect, 7.6. As can be seen from inspection of Table 4.15, which details the results of the ANOVA, this difference was significant [$F(1,8)=41.32$, $p < .001$]; Hypothesis XIV is supported. The main effect for experience was not significant nor was there a significant Experience x Judgment interaction-effect.

Ease-of-Interpretability of Individual Physiological Components

Ease-of-interpretability ratings.--For both groups of evaluators for all record sets, mean ease-of-interpretability ratings were highest for respiration, followed by cardiovascular activity and GSR, respectively. These data are shown in Table 4.16. This result is not consistent with

the results reported by Kubis in a study of experimental lie-detection where such ratings were highest for GSR, cardiovascular activity, and respiration, in order.[11]

TABLE 4.15.--Analysis of Variance Table for Mean Total Ease-Of-Interpretability Ratings on Correct and Incorrect Judgments.

Source	SS	df	MS	F	p<
A (A=Experience-Low, high)	2.31	1	2.31	1.06	.25
E (E=Evaluators):A	17.45	8	2.18		
J (Judgments-correct, incorrect)	1.25	1	1.25	41.32	.0003
A X J	.02	1	.02	.59	-
J X E:A	.24	8	.03		
TOTAL	21.27	19			

TABLE 4.16.--Mean Ease-Of-Interpretability Ratings of The Three Physiological Components on All Record Sets.

Evaluator Experience Level	Physiological Component		
	Respiration	GSR	Cardio
Low	2.87	2.30	2.40
High	3.00	2.55	2.75
Combined	2.93	2.43	2.57

[11] J. Kubis, Studies in Lie Detection: Computer Feasibility Considerations, Tech. Report 62-205 (Arlington, VA.: Armed Services Technical Information Agency, June, 1962), prepared for Air Force Systems Command, Contract No. AF 30 (602)-22700, Project No. SS34, Fordham University, 1962, 70.

Treating each evaluator's mean-rating for correct and incorrect judgments for each component as a dependent variable, t-tests for correlated means ($t_{dep.}$) were conducted. Since no predictions were made concerning differences between the two judgments for individual components, two-tailed tests were used and although mean ratings for both experience-levels of evaluators are presented in Table 4.17, the levels were not treated as a factor in the analysis.

As shown in Table 4.17, the mean ease-of-interpretability ratings were higher in correct than incorrect judgments for both respiration, 3.11 and 2.74, and cardiovascular activity, 2.64 and 2.47; these differences were significant ($p < .001$; $t_{dep.}=6.4$ and 7.3, respectively). For GSR the mean rating of correct was very slightly lower than of incorrect judgments and not significant ($t_{dep.}=-.25$, $p > .10$).

TABLE 4.17.--Mean Ease-Of-Interpretability Ratings of The Three Physiological Components on Correct and Incorrect Judgments.

	Physiological Component					
Evaluator Experience Level	Respiration		GSR		Cardio	
	Cor.	Incor.	Cor.	Incor.	Cor.	Incor.
Low	3.07	2.73	2.25	2.32	2.46	2.29
High	3.15	2.74	2.57	2.52	2.81	2.64
Combined	3.11	2.74	2.41	2.42	2.64	2.47

Ease-of-interpretability scores.--While no predictions were made concerning the ease-of-interpretability scores of individual physiological components, a limited discussion of these results will be undertaken here. Separate four-way analysis of variance, repeated measures, was conducted, treating as dependent variables the ease-of-interpretability scores for the three components. The four factors were identical to those discussed in previous sections of this chapter dealing with such analysis: Experience, Verification, Truthfulness, and Crime-type, the latter three treated as repeated measures. The results of these three analyses are discussed below.[12]

(1) Respiration.--The ANOVA conducted on the respiration ease-of-interpretability scores revealed no significant main effects for the experience or crime-type factors. However, respiration was judged easier to interpret on record sets in the verified than the unverified category [$F(1,8)=40.87$, $p< .001$], and easier in the deceptive than the truthful category [$F(1,8)=102.65$, $p< .001$]. The interpretation of these main effects, however, is complicated by interaction-effects which emerged from the analysis. The mean respiration "ease" scores pertaining to two of these

[12] ANOVA tables for ease-of-interpretability scores for respiration, GSR, and cardiovascular activity are displayed in Appendix E, Tables E.4, E.5, and E.6, respectively.

interactions are shown in Figures 4.5 and 4.6, which, generally, indicate that respiration was judged easier to interpret on record sets in the crime-against-a-person category than in the property-crime category, and easier on record sets in the deceptive category than in the truthful, across the levels of the Verification factor, respectively.

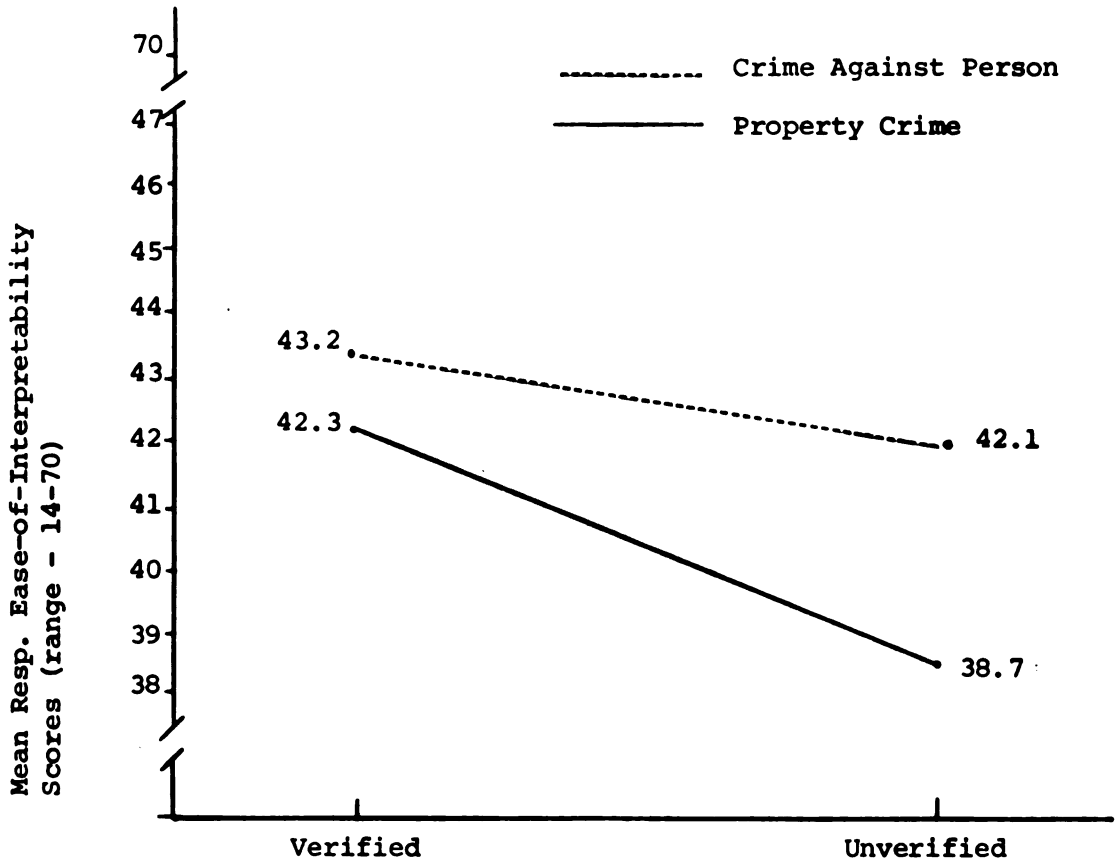

Figure 4.5.--Mean respiration ease-of-interpretability scores for record sets in both crime classifications for the verified and unverified conditions.

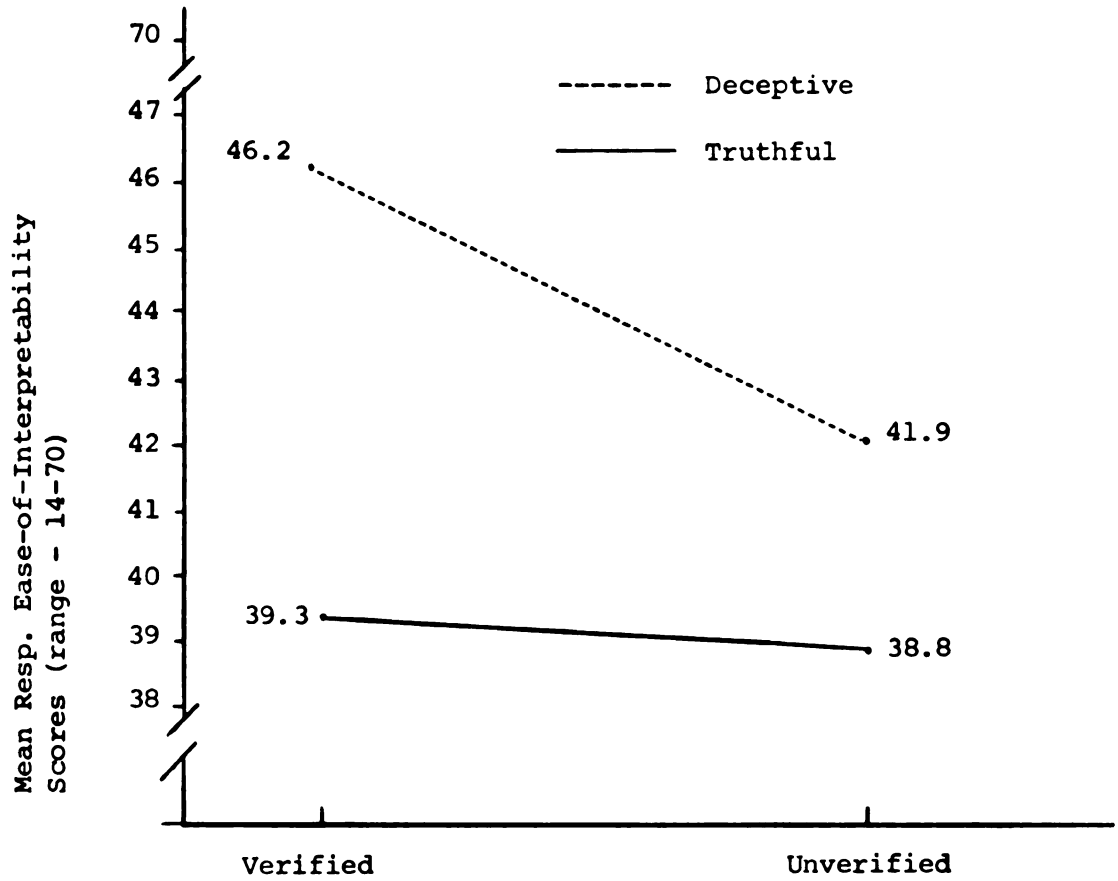

Figure 4.6.--Mean respiration ease-of-interpretability scores on record sets in the truthful and deceptive categories for the verified and unverified conditions.

Two other significant interaction effects which emerged from the analysis are shown in Figures 4.7 and 4.8. As can be seen from inspection of these figures the interactions are disordinal in nature; in spite of this it can be stated that respiration was judged essentially easier to interpret for both crime types in the deceptive category than in the truthful category (Figure 4.7), and for both groups of evaluators in the verified condition than in the unverified (Figure 4.8).

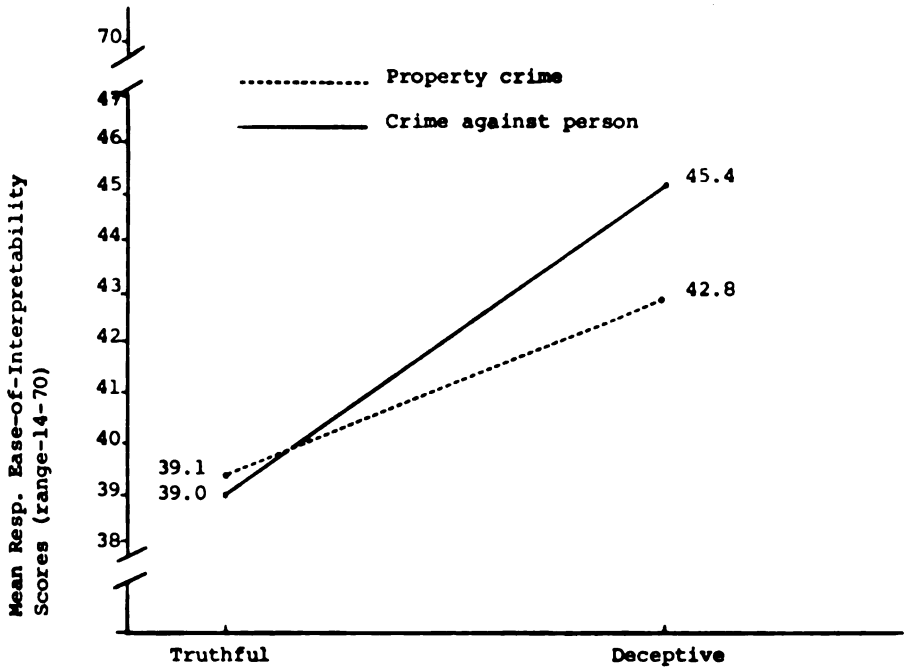

Figure 4.7--Mean respiration ease-of-interpretability scores for record sets in the two crime classifications in the truthful and deceptive categories.

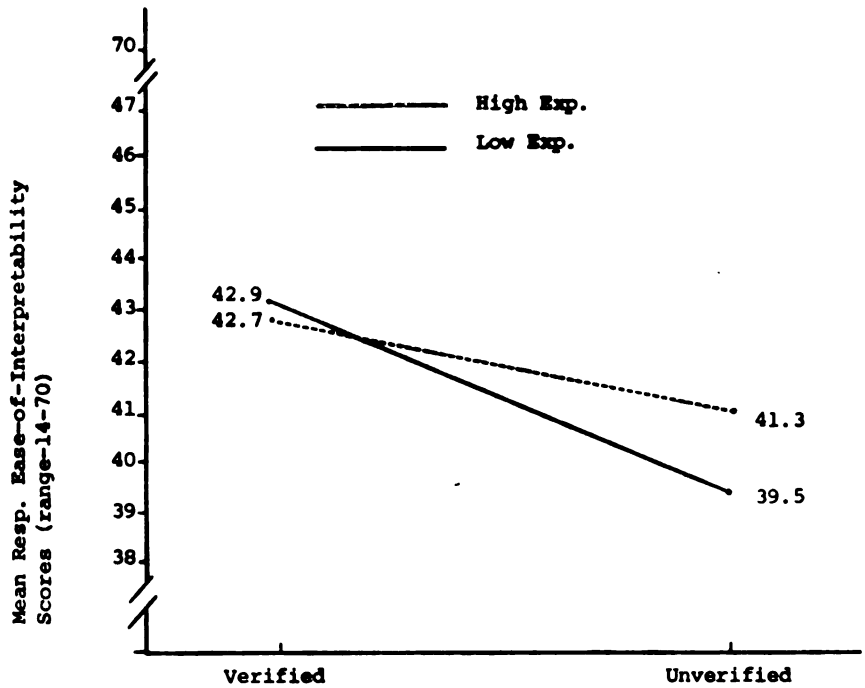

Figure 4.8.--Mean respiration ease-of-interpretability scores for high and low experience evaluators on record sets in the verified and unverified conditions.

(2) GSR.--No significant main effects were apparent from the ANOVA conducted on the ease-of-interpretability scores for GSR. Two significant interaction effects, however, were apparent, a Verification x Truthfulness effect [F (1,8)=12.13, p< .008], and a Verification x Truthfulness x Crime type effect [F (1,8)=6.29, p< .04]. The mean scores pertaining to the first of these interactions are shown in Figure 4.9, for the second in Figure 4.10. However, the meaning of these interactions is too obscure to be discussed here.

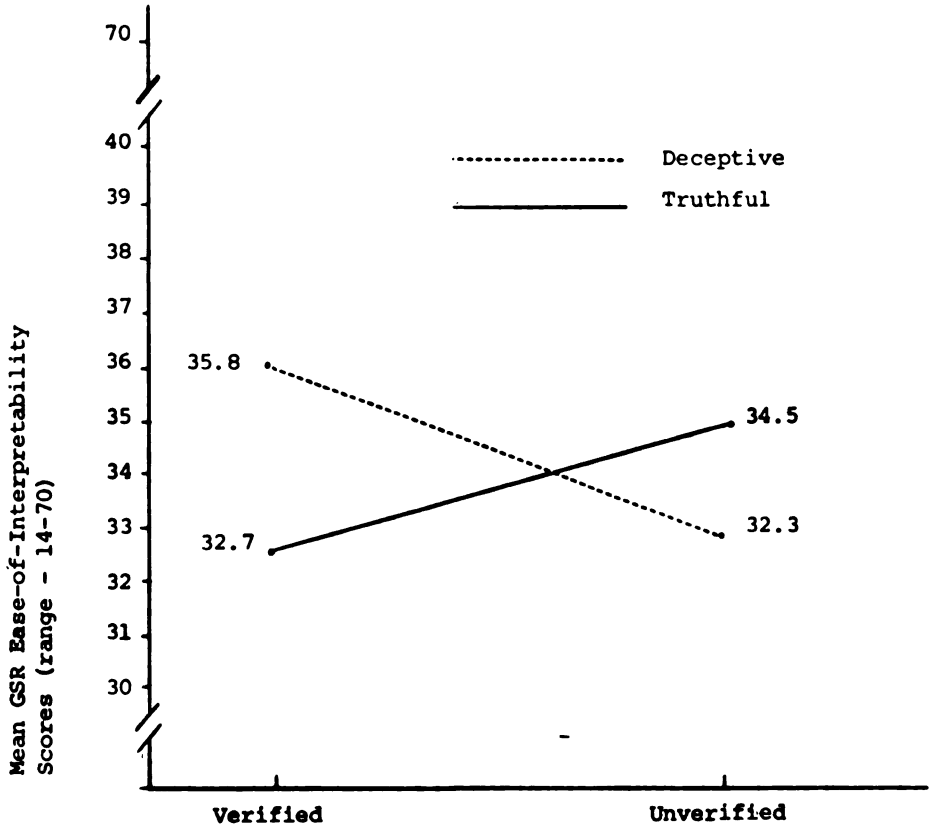

Figure 4.9.--Mean GSR ease-of-interpretability scores on record sets in the truthful and deceptive categories for the verified and unverified conditions.

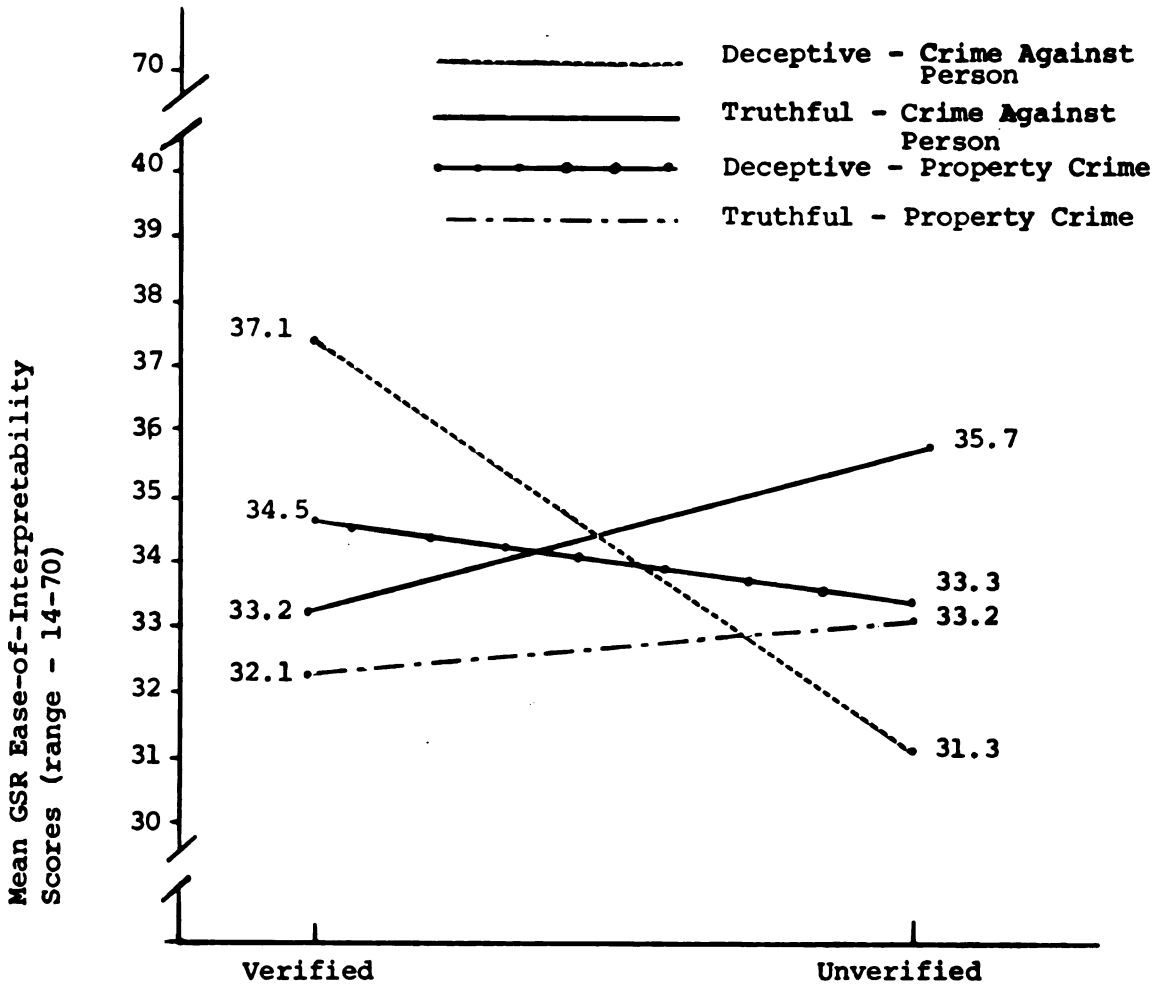

Figure 4.10.--Mean GSR ease-of-interpretability scores on deceptive and truthful crime against a person and property crime record sets for the verified and unverified conditions.

(3) Cardio.--The only significant effects which were apparent from the ANOVA conducted on the cardio "ease" scores concerned the main effects for the Truthfulness and Crime-type factors. Examination of the mean scores for these effects shows that cardiovascular activity was judged easier to interpret on record sets in the deceptive than in the truthful category, with means of 37.7 and 34.3, respectively

[F (1,8)=59.27, p< .001]; and easier for record sets in the crime-against-a-person than the property-crime category, with means of 37.2 and 34.8, respectively [F (1,8)=9.87, p< .01].

Numerical Evaluation

Forty of the 112 record sets, five from each of the eight categories from which sets were drawn, were numerically scored by evaluators; scores for each of the three physiological components in each set had a theoretical range from plus 24 to minus 24; a combined score (the sum of the scores for the three components) had a theoretical range from plus 72 to minus 72. In all cases positive scores indicated greater responsiveness to control-questions in a record set, i.e., truthfulness; negative scores, greater responsiveness to relevant questions, i.e., deception.

Accuracy

The accuracy-scores discussed previously in this chapter were not independent of evaluators' numerical scores; comparisons are thus inappropriate. However, because numerical evaluation provides a mean of assessing the relative accuracy of the individual physiological components, a brief description of the accuracy of such evaluation follows.

Only seven of the ten evaluators, four in the low-experience group, three in the high-experience group, scored the record sets assigned to numerical evaluation. To determine the accuracy of these evaluators' judgments as based

solely on numerical scores, a procedure reported by Barland was used.[13] For combined scores a decision-rule which categorized as "inconclusive" all scores from plus to minus four was applied; that is, for scores on record sets in the truthful category (of which there were twenty assigned to numerical evaluation) any combined score greater than plus four was correct, less than minus four, incorrect, between plus and minus four, inconclusive. For record sets in the deceptive category, the reverse of this procedure determined correct and incorrect judgments. For the scores of individual components the decision rule used determined as inconclusive all scores from plus to minus one, inclusive.

Table 4.18 displays the average accuracy obtained when the decision rules discussed above were applied to the scores of all evaluators. For combined scores, 42 percent were correct, 32 percent incorrect, and 26 percent inconclusive. For individual components cardio scores were the most accurate and GSR the least accurate, at 44 and 37 percent, respectively. It should also be noted that the scores for GSR were inconclusive almost twice as often as those for the other two components.

---

[13] G. Barland, "The Reliability of Polygraph Chart Evaluations" (paper presented at American Polygraph Association Seminar, August 15, 1972, Chicago, Illinois).

TABLE 4.18.--Average Percent Accuracy of Evaluators' Judgments Based on Numerical Scores.

	Judgments		
Component	Correct	Incorrect	Inconclusive*
Respiration	43%	37%	20%
GSR	37%	24%	39%
Cardio	44%	37%	19%
Combined	42%	32%	26%

*The boundaries of the inconclusive region were ± 1, inclusive, for each individual component and ± 4, inclusive, for the score for all components combined.

When inconclusive judgments are eliminated, the average accuracy of all seven evaluators was 57 percent for combined scores, 53, 54, and 60 percent for respiration, cardio, and GSR scores, respectively. These data are displayed in Table 4.19, which also shows the accuracy of individual evaluators, excluding inconclusive scores.

TABLE 4.19.--Percent Accuracy of Individual Evaluators Based on Numerical Scores (Excluding Inconclusives*).

Evaluator	1	2	3	4	5	6	7	Mean
Component								
Respiration	57	47	53	52	62	53	51	53
GSR	63	61	59	65	61	63	47	60
Cardio	48	62	51	57	53	58	50	54
Combined	59	55	56	58	62	55	56	57

*The boundaries of the inconclusive region were ± 1, inclusive, for each individual component and ± 4, inclusive, for the score for all components combined.

A further analysis of the accuracy of judgments based on numerical scores is shown in Table 4.20, which compares the average accuracy for record sets in the verified to that in the unverified category, excluding inconclusives. In the former category, GSR scores were correct an average of 63 percent; this was higher than the accuracy of the other two individual components and of the combined scores. For record sets in the unverified category combined scores were more accurate at 61 percent, than those of the individual components; respiration scores were slightly more accurate, at 59 percent, than either GSR or cardio scores.

TABLE 4.20.--Average Percent Accuracy on Verified and Unverified Record Sets Based on Numerical Scores (Excluding Inconclusives*).

	Category	
Component	Verified	Unverified
Respiration	47%	59%
GSR	63%	58%
Cardio	50%	58%
Combined	53%	61%

*The boundaries of the inconclusive region were $\pm$ 1, inclusive, for each individual component and $\pm$ 4, inclusive, for the score for all components combined.

Reliability

To determine the reliability, i.e., the extent of inter-evaluator agreement, of the scores derived from numerical evaluation, Pearson product-moment correlation coefficients (r) were computed for the set of scores for each of the possible pairs of evaluators. Since there were seven evaluators, twenty-one pairings were possible; correlations were calculated for each of these pairs for respiration, GSR, cardiovascular, and combined scores.

Table 4.21 displays the correlation matrix for the pairs of evaluators with respect to combined scores. As indicated, these correlations ranged from .45 to .82; the mean was .65.[14] Tables 4.22, 4.23, and 4.24 display the correlations obtained for respiration, GSR, and cardio scores, respectively. The range for respiration scores was from .35 to .82, with a mean of .60; for GSR, from .61 to .86, with a mean of .74; and for cardio, from .33 to .78, with a mean of .60. Thus, there was greater agreement between evaluators on GSR scores than on either of the other two components or on combined scores.

To clarify the reliability of numerical scoring, correlations were calculated using the scores for the pairings of evaluators on the record sets in the verified and unverified categories separately. The complete correlation

---

[14] All mean correlations were calculated using the r-Z transformation on raw correlation coefficients.

TABLE 4.21.--Correlations of Combined Scores.

Evaluator	1	2	3	4	5	6	7
1		.70	.51	.72	.62	.62	.73
2			.52	.76	.59	.71	.64
3				.45	.58	.46	.68
4					.74	.69	.77
5						.60	.82
6							.60

TABLE 4.22.--Correlations of Respiration Scores.

Evaluator	1	2	3	4	5	6	7
1		.65	.40	.71	.46	.61	.65
2			.35	.70	.45	.65	.61
3				.45	.56	.35	.66
4					.62	.66	.71
5						.51	.82
6							.65

TABLE 4.23.--Correlations of GSR Scores.

Evaluator	1	2	3	4	5	6	7
1		.81	.66	.82	.83	.86	.79
2			.67	.73	.68	.83	.66
3				.71	.65	.67	.61
4					.78	.82	.71
5						.79	.67
6							.67

TABLE 4.24.--Correlations of Cardio Scores.

Evaluator	1	2	3	4	5	6	7
1		.62	.49	.64	.63	.40	.74
2			.53	.73	.62	.63	.61
3				.33	.44	.41	.57
4					.77	.58	.74
5						.59	.78
6							.55

matrices for these data are displayed in Appendix F, Tables F.1 through F.8. The mean correlations for these data, however, are displayed in Table 4.25; as shown, in all cases the mean correlations were higher for record sets in the unverified than in the verified category, although none of the differences were significant when tested by a t-test for correlated means ($t_{dep.}$)[15]

TABLE 4.25.--Comparison of Mean Correlations of Numerical Scores of Verified to Unverified Record Sets.

	Record Sets		
Score	Verified $\bar{r}$	Unverified $\bar{r}$	t dep.
Cardio	.60	.65	-1.85*
GSR	.73	.74	- .63**
Respiration	.60	.65	- .92**
TOTAL	.64	.70	-1.65**

* p< .10
** p> .10

[15]The t-tests were calculated in all cases by transforming the correlations to Z-variables; these variables were then used as the dependent measure. Since no predictions were made, two-tailed tests were used.

Chapter V

DISCUSSION

The results of this study essentially indicate the following: (1) That depending solely on polygraphic recordings obtained from field examinations conducted by control-question technique, the judgments of trained evaluators are accurate well beyond chance levels. (2) That there is substantial agreement (reliability) among evaluators concerning truth/deception judgments made on polygraphic recordings. (3) That the nature of polygraphic recordings -- the categories from which they are drawn -- is a more important variable in blind analysis than is the experience of evaluators.

Accuracy of Judgments

That more experienced evaluators in this study were not significantly more accurate in their judgments than those less experienced is generally contrary to the findings of previous researchers. A plausible explanation for this difference lies in the definition of "experience". Horvath and Reid, for instance, found that incompletely trained evaluators with less than six months' experience were less accurate than those fully trained and with varying degrees

of active experience. In the present study, however, all evaluators had completed a formal training course, and although some were still interns, all had a minimum of eight months' active experience. It is reasonable to suspect, therefore, that given evaluators of a minimum level of experience, the nature of recordings is more critical in blind analysis than is experience per se.

The effect of the specific sources of the polygraphic recordings on accuracy is apparent in analysis of the interaction-effect pertaining to accuracy scores, as shown in Figure 4.2, page 128. In all but one of the eight categories of record sets, accuracy was higher for those of deceptive than of truthful subjects. This finding contrasts with prior research reported by field examiners but is consistent with results reported by Barland in his experimental study of lie-detection.[1] However, it is also apparent that this finding is complex and intricate.

Inspection of Figure 4.2 shows that record sets in the "crime against a person" classification were judged deceptive more often than all others; hence the likelihood of false positives was greatest, and of false negatives least, in this classification, regardless of verification. The most likely explanation of this result is that relevant

---

[1]G. Barland, "An Experimental Study of Field Techniques in Lie Detection," (unpublished Master's Thesis, University of Utah, 1972), 38.

questions pertaining to investigations of crimes against a person elicit stronger physiological responses from both truthful and deceptive subjects than do such questions pertaining to property-crime investigations. In other words, crimes against a person are, by nature, more emotionally weighted, a condition heightening the possibility of false positives in blind analysis of physiological responses.

There is no completely satisfactory explanation for other aspects of the interaction pertaining to accuracy-scores. For instance, it is not clear why accuracy increased from the verified to unverified condition in judgments made on record sets in the "deceptive/crime against a person" and "truthful/property crime" categories when it decreased for other categories of record sets. Nor is it obvious why in the unverified condition record sets of truthful subjects in the property crime classification were more accurately judged than were those of deceptive subjects in the same classification. The latter finding, however, probably reflects the lack of uniform numbers of control question tests in record sets.

Evaluators were more accurate on record sets limited to the basic battery of control-question tests than those including additional tests. Inspection of the distribution of record sets containing only the basic battery indicates six such sets in the "truthful/property crime" and only four in the "deceptive/property crime" category, both in the

unverified condition.[2] Thus, it is possible that higher accuracy in the former category was due to the predominance of more accurately judged record sets. In spite of this possibility, however, it is notable that other results pertaining to accuracy-scores are not explained by differences in the number of control-question tests in record sets.

That evaluators were more accurate on record sets with less rather than more physiological data (CQ tests) conflicts with Rouke's results. Rouke reported greater accuracy and reliability when evaluators of experimentally-derived lie-detector (GSR) recordings were given additional data.[3] It seems likely that in the present study record sets containing only the basic battery of CQ tests were clearer in their indications of truthfulness and deception than those including additional tests. This explanation suggests that the examiners who actually conducted the testing supplemented it with additional tests when the basic battery was ambiguous in its indications, that additional tests and "stimulation" strategies may not clarify response-data to the extent which field examiners contend.

Barland, in experimental lie-detection, reported that when he combined the numerical scores of a group of evaluators, the combined "average scores" were more accurate than the

---

[2]See Table 3.2, page 100.

[3]F. Rouke, "Evaluation of The Indices of Deception in the Psychogalvanic Technique" (unpublished Ph.D. dissertation, Fordham University, 1941), 46-47.

average accuracy of individual evaluators, that pooling individual decisions increased accuracy.[4] His results are generally supported by the present study's findings pertaining to collective accuracy: the greater the number of evaluators in agreement, the greater the accuracy.

The nature of the criteria for assessing accuracy in this study was such that accuracy-scores were clearly unrelated to the validity of lie-detection in the field. The requisite criteria of judgments made on verified record sets were confessions; thus, within reasonable limits, "ground truth" against which evaluators' judgments could be compared was known. As Orne has argued, however, such judgments reflect only the extent to which evaluators can reliably identify those aspects of physiological data which they view as indicative of truthfulness and deception.[5] In other words, the examiners' actual judgments in all verified situations were correct (valid); it is not known if the examiners relied on physiological or other information to make such judgments.

On the other hand, the criteria of accuracy on unverified record sets were the judgments of the testing

---

[4] G. Barland, "The Reliability of Polygraph Chart Evaluations," (paper presented at American Polygraph Association Seminar, August 15, 1972, Chicago, Illinois), 7.

[5] M. Orne, "Implications of Laboratory Research for the Detection of Deception," Polygraph, 2 (1973), 179.

examiners. Of course, under such conditions neither "ground truth" nor the nature of the information which the testing examiners used to make such judgments was known. Accuracy-scores, then, whether on verified or unverified record sets, are essentially measures of reliability, agreement between examiners' judgments based on many sources of information, e.g., physiological data, behavioral characteristics of subjects, investigators' reports, etc., and evaluators' judgments based solely on physiological data.

In view of the above argument it is noteworthy that accuracy-scores, while generally "correct" well beyond chance levels overall, were not substantially higher on verified than on unverified record sets. This result suggests that polygraphic recordings themselves are relatively stable from the first situation to the second. It also suggests that while physiological data are a substantial contribution to (police) examiners' judgments in actual field-testing, other sources of information probably have a considerable influence. In other words, as many field-examiners contend, lie-detection in the field is a diagnostic technique the validity of which is neither completely determined by, nor independent of, physiological information.

Reliability of Judgments

The consistency of evaluators' judgments in this study substantiates prior research, whether experimental or field-based, that there is considerable agreement among independent

evaluators as to the criteria believed associated with deception. That is, that blind analysis of polygraphic recordings by trained evaluators is an objective, reliable, procedure.

Pairs of evaluators in this study agreed on an average of 69 percent of their judgments. Barland reported an average agreement of 95.5 percent for (pairs of) six field trained evaluators of experimentally-derived polygraphic recordings.[6] The difference in these results may be due to the nature of the polygraphic recordings, i.e., experimental as opposed to field. On the other hand, it is also likely that the difference is partially explained by the fact that the evaluators in Barland's study scored the recordings numerically. The percentage of agreements reported represents the percentage of incidence of paired evaluators' scores indicating a definite decision; thus, disagreements caused by one of a pair's scores falling into the inconclusive region were not counted.

It was apparent in this study that evaluator reliability did not substantially vary whether judgments made on verified or unverified record sets were considered; for both categories reliability coefficients were quite high, .89 and .85, respectively. This result supports the earlier suggestion

---

[6]Barland, "The Reliability of Polygraph Chart Evaluations," op. cit., 5.

in this chapter that there is a high degree of consistency in polygraphic recordings, whether derived from verified or unverified investigations.

Although the accuracy and reliability of the judgments made by the evaluators in this study were quite substantial, it is apparent that these results were not as convincing as those reported in other somewhat similar studies dealing with field-derived polygraphic recordings. Horvath and Reid, for instance, reported an average accuracy of 87.7 percent for ten evaluators' judgments on the polygraphic recordings of forty subjects. A similar figure, 86 percent, was reported by Hunter and Ash for seven evaluators' judgments on twenty polygraphic recordings. In both of these studies errors were almost identically balanced; that is, false positives occurred nearly as often as false negatives.

Some of the possible explanations for such inconsistencies between prior research and the present study are quickly eliminated as unlikely; others appear more relevant. Of probably minimal influence on differential results are the following:

1) In previous sections of this study it is suggested that verified recordings may be more accurately interpreted than those which are unverified, implying that the Horvath/Reid and Hunter/Ash studies, using only verified recordings, biased results in favor of higher accuracy. However, evaluators

in the present study were not substantially more accurate on verified than on unverified recordings, average accuracy on the record sets in the former category being lower than that reported in previous studies.

2) A second possible explanation is that evaluators in prior studies may have had more experience in, or been more adept at, interpreting polygraphic recordings. The explanation is unconvincing since in this study evaluators, actively engaged in lie-detection for a period of years, were, on the average, less accurate than those in prior studies who had not yet completed a six-month training course.

3) Finally, evaluators in prior studies were not only given polygraphic recordings but were also briefed about the investigations from which the recordings were obtained. While Holmes has demonstrated that accuracy increases when evaluators are given information in addition to recordings[7] it is exceedingly doubtful that the slight information given evaluators in prior studies can account for the substantial increment in accuracy over that in the present study.

The most convincing explanations for the findings in the present study, and certainly factors which make it difficult to draw direct comparisons between this and other research, include the following:

---

[7]W. Holmes, "The Degree of Objectivity in Chart Interpretation," Academy Lecture on Lie Detection, II, V. Leonard (ed.) (Springfield, Illinois: C.C Thomas, 1958), 62-70.

1) In contrast to prior studies, polygraphic recordings in the present study were selected at random from a pre-defined population. While randomization was, for this study, a desirable characteristic, it eliminates the possibility of control for any influence of examiner-subject interaction on polygraphic recordings. In other words, recordings in this study were included without regard for the capabilities of the examiners who had conducted the examinations from which the recordings derived. In fact, it became apparent during the study that some of the recordings were derived from examinations conducted by examiners who were, during the years from which the sample was drawn, interns.

On the other hand, recordings evaluated in prior studies were, in each case, obtained from examinations conducted by the same experienced examiner. Obviously, any effect of examiner-subject interaction on physiological recordings was, at the least, minimized. Said in another way, variability due to differences between examiners was eliminated.

It should be noted here that the lack of any significant differences between experience-levels of evaluators in the present study, does not refute the above considerations. That examiners acting as evaluators apparently do not differ in ability to interpret physiological data is not to say that experience is an unimportant variable in conducting

polygraphic examinations. In fact, in view of Orne's argument that the primary variables in lie-detection are psychological, not physiological, in nature,[8] experience is probably a critical determinant of the outcome of such examinations. In other words, it is experience that probably permits an examiner to adjust more effectively to complex situational demands.

2) Two of the prior studies have dealt with polygraphic recordings of subjects involved in investigations undertaken by private or commercial examiners, whereas in the present study the recordings were of subjects involved in investigations conducted by police agencies. There may be obvious and dramatic differences between the two subject populations in regard to many of the variables known to influence autonomic activity, and, more generally, lie-detection. For instance, variables such as intelligence, ethnicity, age, and generally, personality and psychological make-up, are probably important determinants of response-data obtained during polygraphic examinations.[9] Moreover, as Orne has pointed out, examinations conducted by private

---

[8]Orne, "Implications of Laboratory Research for the Detection of Deception," op. cit., 188.

[9]See: G. Barland and D. Raskin, "The Use of Electrodermal Activity in the Detection of Deception," Pre-publication copy to appear in W. Prokasy and D. Raskin (Eds.), Electrodermal Activity in Psychological Research (New York: Academic Press, in press), 31-39.

examiners may differ from those of police examiners with respect to the motivation of the subject, and the amount and nature of the information available to the examiner prior to the testing.[10] All of these variables, singularly or in combination, might make blind analysis of polygraphic recordings of police examinations more difficult than analysis of those obtained from commercial situations.

3) In examinations conducted by police examiners, the degree of stress on the subject is presumably greater than in those conducted by commercial examiners. Such stress is believed to increase detectability; thus, it could be suggested that evaluators of recordings obtained from police examinations would be more accurate than those who judge recordings obtained under different circumstances. However, neither Holmes's findings[11] nor the results of the present study support such a suggestion. It may be that there is a threshold of stress, encountered primarily in police situations, beyond which the detectability of truthfulness and deception in blind analysis of polygraphic recordings decreases; or, said in another way, beyond which the ambiguity of responses increases. Such ambiguity might also increase false positives.

---

[10] Orne, "Implications of Laboratory Research for the Detection of Deception," op. cit., 188.

[11] Holmes, "The Degree of Objectivity in Chart Interpretations," op. cit., 67.

4) Finally, evaluators in the present study were denied the advantage of some physiological data available to the testing examiners. For methodological reasons, "yes" tests were eliminated from all record sets; it is not clear whether such tests were included in prior research.

Although it is possible that the elimination of "yes" tests decreased overall accuracy, it probably did not affect the relative results. With but one exception "yes" tests had to be eliminated from record sets in the deceptive category; hence, if anything, accuracy would have increased only on these record sets. Judgments of record sets in the truthful category would have been unaffected.

Confidence in Judgments

In general, the results pertaining to confidence-scores are consistent with those of accuracy-scores. More experienced evaluators were not significantly more confident than those less experienced, nor was confidence significantly greater on verified than on unverified record sets. These results lend support to explanations previously advanced in the discussion concerning the accuracy of judgments.

That confidence-scores were significantly higher on the record sets in the deceptive than in the truthful category also supports prior discussion. In blind analysis the physiological responses believed to be associated with deception not only are more accurately but also more confidently

judged than those indicative of truthfulness. This result is consistent irrespective of verification involved.

While field-research dealing with the relationship between confidence-ratings and accuracy has not been reported, there are two experimental studies of lie-detection which have explored this issue. Kubis reported that independent evaluators of polygraphic recordings "had greater confidence in those decisions ultimately verified as correct than they did in those which were incorrect."[12] In a later study, Moroney substantiated Kubis's findings.[13]

The results of the present study clearly support those reported by Kubis and Moroney: confidence was significantly greater on correct than incorrect judgments for both experience-groupings of evaluators. While the practical significance of this finding is unclear, it suggests that the more ambiguous the recordings, the greater the possibility for error in blind analysis, regardless of the experience of the evaluator. When evaluators identified those aspects of physiological data believed to be indicative of truthfulness and deception, confidence increased; when those aspects were

---

[12] J. Kubis, Studies In Lie Detection: Computer Feasibility Considerations, Tech. Report 62-205 (Arlington, VA.: Armed Services Technical Information Agency, June, 1962), prepared for Air Force Systems Command, Contract No. AF 30 (602)-22700, Project No. 5534, Fordham University, 1962, 68.

[13] W. Moroney, "The Detection of Deception as a Function of PGR Methodology," (unpublished Ph.D. dissertation, St. Johns University, 1968, Ann Arbor, Michigan: University Microfilms, 1969, No. 69-7125).

less apparent, confidence decreased. Moreover, it is interesting that this finding obtained even though the criteria for assessing accuracy were not the same for verified and unverified conditions, suggesting again that the nature of the recordings in the two conditions is relatively consistent.

Ease of Interpretability of Record Sets

The results concerning the total ease-of-interpretability scores are both consistent and inconsistent with results pertaining to accuracy and confidence-scores. In regard to consistencies, it is apparent that the experience of evaluators did not significantly influence total "ease" scores. Contrary to Horvath and Reid's suggestion, when in blind analysis, more experienced evaluators apparently do not find it easier than do the less experienced to interpret polygraphic data, "to apply consistently the fine points of the [control question] theory."[14] However, as will be discussed, "ease" scores may not have been a very effective measure of truthfulness/deception indicated by physiological data.

A second finding regarding the total "ease" scores and supporting other findings was that record sets in the deceptive category were judged significantly easier to

---

[14]F. Horvath and J. Reid, "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," J. Crim. Law, Crim. and Pol. Sci., 63 (1972), 281.

interpret than those in the truthful category, whether verified or unverified. This finding, of course, is consistent with the greater confidence and accuracy scores on "deceptive" record sets.

Total "ease" scores decreased considerably from the verified to the unverified condition for record sets in the deceptive category, while for those in the truthful category they increased slightly. (These same effects were also apparent in confidence scores.) Again, an explanation of these results may lie in the lack of uniform numbers of control-question tests in record sets. In the deceptive category it is apparent that there were more record sets containing only the basic battery in the verified than in the unverified condition; for sets in the truthful category the basic battery was apparent more often in the unverified than the verified condition.[15] Thus, the direction of "ease" scores across the levels of verification may reflect merely differences in the number of record sets containing only the basic battery, presumably easier to interpret than other record sets. It is clear, however, that such differences do not account for the relationship between the ease of interpretability of truthful and deceptive record sets; "deceptive" were easier to interpret than "truthful" irrespective of the number of record sets in each of these categories containing only the basic battery.

---

[15]See Table 3.2, page 100.

Total ease-of-interpretability scores were significantly higher in the verified than in the unverified condition. This result seems to conflict with other results, since neither accuracy nor confidence scores were significantly different in these two conditions. It is likely, however, the "ease" scores were not a measure of the degree to which an evaluator could discriminate between control-relevant question responses; hence, they were not directly related to accuracy. The "ease" scores were apparently regarded by evaluators as an index of the general level of the responsiveness of the physiological data in record sets, perhaps irrespective of truthfulness/deception indications. This explanation helps clarify why record sets in the crime-against-a-person category were judged significantly easier to interpret than those in the property-crime category; it is also consistent with the explanation previously advanced concerning the accuracy-score results: Investigations concerning crimes against a person are, by nature, more emotionally weighted than those pertaining to property crimes; thus, the general level of responsiveness for record sets in the former category is greater than in the latter.

Results of analysis of the ease-scores for individual components are essentially similar to those for total "ease"-scores. Respiration and cardio were easier to interpret for record sets in the deceptive category than those in the

truthful. This same general result was also found for GSR "ease" scores but only in the verified condition -- an exception not readily explained.

The mean total ease-of-interpretability ratings were significantly higher on correct than on incorrect judgments. This result approximates Kubis's findings that for independent evaluators correctly judged records are easier to interpret than those incorrectly judged.[16] Other results of the present study, however, are strikingly dissimilar to those reported by Kubis.

In the present study physiological components were rated for ease-of-interpretability in the following order: respiration, cardio, and GSR, the first two components judged significantly easier to interpret on correct than on incorrect judgments. These results, corresponding with anecdotal evidence offered by field examiners concerning the relative merits of the individual components,[17] do not correspond with those of Kubis's laboratory study. Kubis found GSR, cardiovascular, and respiratory activity, in that order, easier to interpret and found the ratings for all three components higher for correct than incorrect decisions.[18] There are several explanations for these differences.

---

[16]Kubis, Studies in Lie Detection: Computer Feasibility Considerations, op. cit., 70-71.

[17]J. Reid and F. Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique (Baltimore: Williams and Wilkins, 1966), 40.

[18]Kubis, Studies in Lie Detection: Computer Feasibility Considerations, op. cit., 70.

Field examiners contend that for their purposes GSR is less useful as an indicator of deception than are respiration or cardiovascular activity. Thus, since the present study involved field polygraphic data evaluated by field-trained evaluators, "ease" ratings may be reflecting the particular orientation of these evaluators. Comparison of the simplicity of GSR responses to the complexity of respiratory and cardiovascular responses, however, detracts from this explanation.

A second explanation of the differences may be that in the field the level of subject affect, being higher than in laboratory situations, distorts GSR responses to the extent that they are, in fact, more difficult to interpret than are respiratory or cardiovascular responses. This explanation is consistent with the claims of field examiners,[19] although there is some indication that such claims may not be legitimate.[20]

Finally, differences in instrumentation in Kubis's laboratory situation and the typical field situation may affect GSR responses. Laboratory equipment such as that used by Kubis is usually more sophisticated than field equipment. Moreover, in field situations the apparatus for recording

---

[19] Reid and Inbau, Truth and Deception: The Polygraph ("Lie Detector") Technique, op. cit., 220.

[20] Barland, "An Experimental Study of Field Techniques in Lie Detection," op. cit., 50.

cardiovascular activity usually causes some discomfort to the subject. Kubis, however, recorded cardiovascular activity in a manner which precluded discomfort. Thus GSR responses in Kubis's study were uninfluenced by this additional factor whereas such responses as evaluated in the present study may have been degraded.[21] These assumptions concerning the effect of instrumentation differences on GSR responses, however, are not fully supported by evidence reported by Barland,[22] Kugelmass,[23] and Orne.[24]

Numerical Evaluation

Of secondary but real interest here is the accuracy of numerical scores of evaluators. When the scores for all record sets were considered, the evaluators' GSR scores, not counting inconclusives, were more accurate, 60 percent, than those for the other components; while this same result did not obtain when the accuracy of scores was calculated separately for record sets in the unverified condition, GSR

---

[21] Alternate explanations for differences between laboratory and field situations with respect to GSR responses are also possible; see: Barland and Raskin, "The Use of Electrodermal Activity in the Detection of Deception," op. cit., 30-44.

[22] Barland, "An Experimental Study of Field Techniques in Lie Detection," op. cit., 44.

[23] S. Kugelmass, I. Lieblich, A. Ben Ishai, A. Opatowski, and M. Kaplan, "Experimental Evaluation of Galvanic Skin Response and Blood Pressure Change Indices During Criminal Interrogation," J. Crim. Law, Crim. and Pol. Sci., 59 (1968), 623-635.

[24] Orne, "Implications of Laboratory Research for the Detection of Deception," op. cit., 196.

scores were not substantially less accurate than those of the other components. These results are not consistent with the claims of many field examiners that in the field GSR is of relatively little merit compared to other physiological indices of deception. The results are, however, consistent with the results of many experimental lie-detection studies, and they agree with Barland's tentative findings in the field.[25]

One reason for the apparent lack of faith which field examiners have in GSR responses may be that in many situations such responses are too ambiguous, i.e., too labile, however otherwise useful they are. This ambiguity is apparent upon examination of the accuracy of the scores for individual components when "inconclusive" scores are not eliminated. The scores for GSR fell into the "inconclusive" region nearly twice as often as those of the other two components, making GSR scores the least accurate. The ambiguity of GSR responses is also apparent from an inspection of the ease-of-interpretability ratings of individual components; GSR was rated the most difficult of the three components to interpret. It should be noted, however, that the ambiguity of GSR in the field may not be situational in nature, but rather due to the inattentiveness of examiners to instrumentation maintenance or adjustment.

---

[25]Barland, "An Experimental Study of Field Techniques in Lie Detection," op. cit., 50.

Results pertaining to the reliability of numerical scores indicate greater agreement between evaluators on GSR scores than on either of the other two components or on combined scores. With but one exception these results are consistent with Barland's findings concerning relative reliability of evaluators' scores.[26] The exception is that in the present study evaluators did not differ in their consistent scoring of respiratory or cardiovascular responses. In Barland's study, on the other hand, respiratory responses were scored with considerably less consistency than either GSR or cardiovascular responses. It is not clear if this difference in results was due to differences in the nature of the polygraphic recordings used (field as opposed to experimental) or to evaluator differences. However, the former explanation seems more likely since the evaluators in both studies were field-trained.

The consistency of evaluators' numerical scores is surprisingly high, especially since evaluators received only minimal instruction in numerical evaluation, and since such scores reflect primarily relevant/control-question response differences rather than overall judgments of truthfulness/deception. This result further indicates that analysis of polygraphic data by field-trained evaluators is relatively objective and reliable.

---

[26]Barland, "The Reliability of Polygraph Chart Evaluations," op. cit., 5.

Summary

It is clear that in general the results of this study support prior research, that the "blind" judgments of trained evaluators made on field-derived polygraphic recordings are accurate well beyond chance levels and that there is a substantial degree of reliability and objectivity in these judgments. Nevertheless, the results also suggest that it may be inappropriate to talk about the accuracy of blind analysis without first specifying the nature of the investigation from which recordings are drawn, whether for law enforcement or commercial purposes.

The most consistent finding in this study was that pertaining to differences between polygraphic recordings of truthful and deceptive subjects. Not only were recordings of deceptive subjects judged more accurately and confidently, but they were easier to interpret than those of truthful subjects. While it is tempting to apply this result to the general field-situation it is inappropriate to do so. The results of this study pertain only to judgments made by blind analysis, which, as already pointed out, differs substantially from the manner in which judgments are made by examiners in field-settings. It is clear that extensive research is warranted to determine the influence which differential sources of information have on examiners' judgments generally, and on the nature of errors in field lie-detection specifically.

APPENDICES

APPENDIX A

NUMBER OF FOLDERS ASSIGNED TO STRATIFICATION LEVELS

TABLE A.1.--Number of Folders Assigned to Stratification Levels.

Verified			
Truthful		Deceptive	
Crimes Against A Person	Property Crimes	Crimes Against A Person	Property Crimes
47	33	187	213

Unverified			
Truthful		Deceptive	
Crimes Against A Person	Property Crimes	Crimes Against A Person	Property Crimes
311	450	100	105

APPENDIX B

INSTRUCTIONS TO EVALUATORS

General Instructions to Evaluators

Enclosed are the polygraph recordings of 28 subjects in PACKET _____. Would you please analyze each set of recordings and for each subject complete fully the EVALUATOR ANSWER SHEET. PLEASE be sure that you have answered all questions on each sheet for each subject.

Some subjects' recordings are given a number followed by the letters QC. These recordings are to be analyzed according to directions for completing the EVALUATOR ANSWER SHEET and the NUMERICAL EVALUATION SCORE SHEET, as explained on February 8, 1974. In other words, for all subjects complete an EVALUATOR ANSWER SHEET; for subjects whose numbers are followed by a QC complete an EVALUATOR ANSWER SHEET and a NUMERICAL EVALUATION SCORE SHEET.

When you have completed an EVALUATOR ANSWER SHEET (and the NUMERICAL EVALUATION SCORE SHEET, where appropriate), place them in the PACKET envelope along with all of the polygraph recordings. (PLEASE BE CAREFUL NOT TO LOSE OR MISPLACE ANY OF THE RECORDINGS).

You will have one week to evaluate all recordings in any one PACKET. If you finish before this time limit please notify (The Chief Examiner) or me and tell us which PACKET you have completed. DO NOT give the recordings to any other examiner.

NOTE: Valid results depend upon each examiner making his own analysis. So please do not consult with anyone else when making your decisions or discuss your results with any other examiner. If you have any questions concerning the study or the procedure please call before you start your analysis.

THANK YOU.

Instructions For Numerical Evaluation

1. Review each measure (resp., GSR, Cardio) separately in Test I.

2. Compare response in each measure to each of the four relevant questions (consider only questions 3k, 5, 8, and 9) to the response on appropriate Control Questions. (See the Numerical Evaluation Score Sheet to decide which Control Question to consider).

3. Decide if the response to the relevant question is greater or less than the response to the Control Question. If the response to the relevant question is greater the score for that question in the measure you are analyzing could be -1, -2, or -3; depending upon how much greater you believe the response is. For instance, if you are evaluating the respiration measure and the response at question #5 is very much greater than the response at Control Question #6, then you would indicate on the score sheet a -3; if the response is only somewhat greater to the relevant question, then you would score a -2, etc. If there is no difference between the relevant question response and the Control Question response, then you would mark a 0 on the score sheet. On the other hand, if the Control Question response is greater than the response to the particular relevant question you are evaluating then you would mark a +1, +2, or a +3, once again depending upon how much greater you believe the Control Question response to be.

4. Carry out step 3 for each of the four relevant questions and for each measure on TEST I.

5. Repeat steps #3 and #4 for TEST III (following the "number" test).

6. If there are two respiration measures recorded, evaluate ONLY the recording of the lower pneumo; that is, the recording which is nearest the bottom of the chart.

7. You do not have to total your scores, unless you want to, since your decision regarding the subject's truthfulness or deception will already be indicated on your EVALUATOR ANSWER SHEET.

APPENDIX C

SPECIMEN COPIES OF EVALUATOR ANSWER SHEETS

EVALUATOR ANSWER SHEET

DATE:_____ PACKET #_____

EVALUATOR
NAME:_____ RECORD #_____

I. BASED UPON YOUR ANALYSIS OF THE SUBJECT'S RECORDS WOULD YOU CONCLUDE THAT HE IS: (Please circle appropriate number).

A truth-teller	(NDI)	1
A liar	(DI)	2
Inconclusive	(INC)	3

II. WOULD YOU PLEASE RATE THE DEGREE OF CONFIDENCE YOU HAVE IN YOUR ANALYSIS:

No confidence	1
Very doubtful	2
More doubtful than confident	3
More confident than doubtful	4
Very confident	5
Almost certain	6

III. OVERALL, HOW EASY WAS IT TO INTERPRET THESE RECORDS?

Easy to interpret?	Resp.	GSR	Cardio
Very easy	5	5	5
Easy	4	4	4
Average	3	3	3
Difficult	2	2	2
Very Difficult	1	1	1

NUMERICAL EVALUATION SCORE SHEET

TEST I	Q3k-6	Q5-6	Q8-6	Q9-10	Component Total	
PNEUMO						
GALVO						
CARDIO						TOTAL
SUB-TOTAL						

TEST III	Q3k-6	Q5-6	Q8-6	Q9-10	Component Total	
PNEUMO						
GALVO						
CARDIO						TOTAL
SUB-TOTAL						

SPOT TOTALS				

GRAND TOTAL	

SUBJECT # _____

PACKET # _____

EXAMINER _____

APPENDIX D

RESULTS OF INDIVIDUAL EVALUATORS' JUDGMENTS

TABLE D.1.--Distribution of Truth-Deception Judgments of Individual Evaluators.

	Categories of Record Sets																							
	Verified								Unverified															
	Truthful						Deceptive						Truthful						Deceptive					
Evaluators	Crime Against A Person			Property Crime			Crime Against A Person			Property Crime			Crime Against A Person			Property Crime			Crime Against A Person			Property Crime			TOTALS		
Low Experience																											
E_1	9	5	(0)	9	5	(0)	11	3	(0)	8	6	(0)	9	5	(0)	10	4	(0)	10	4	(0)	6	8	(0)	72	40	(0)
E_2	5	9	(0)	5	9	(0)	13	1	(0)	14	0	(0)	3	11	(0)	5	9	(0)	13	1	(0)	11	3	(0)	69	43	(0)
E_3	7	7	(0)	5	9	(0)	11	3	(0)	12	2	(0)	3	11	(0)	9	5	(0)	13	1	(0)	11	3	(1)	71	41	(1)
E_4	5	9	(0)	6	8	(0)	13	1	(0)	11	3	(0)	2	12	(0)	10	4	(0)	10	4	(0)	12	2	(0)	69	43	(0)
E_5	7	7	(0)	9	5	(0)	11	3	(0)	11	3	(0)	5	9	(0)	9	5	(0)	12	2	(0)	6	8	(0)	70	42	(0)
Sub-Total	33	37	(0)	34	36	(0)	59	11	(0)	56	14	(0)	22	48	(0)	43	27	(0)	58	12	(0)	46	24	(1)	351	209	(1)
High Experience																											
E_6	6	8	(0)	8	6	(0)	13	1	(0)	12	2	(1)	8	6	(0)	10	4	(0)	11	3	(0)	10	4	(0)	78	34	(1)
E_7	8	6	(0)	5	9	(0)	13	1	(0)	10	4	(0)	4	10	(0)	8	6	(0)	12	2	(0)	12	2	(0)	72	40	(0)
E_8	9	5	(1)	9	5	(0)	10	4	(1)	10	4	(0)	5	9	(0)	11	3	(0)	11	3	(0)	7	7	(0)	72	40	(2)
E_9	8	6	(2)	10	4	(1)	4	10	(2)	7	7	(2)	6	8	(1)	10	4	(0)	9	5	(0)	6	8	(0)	60	52	(8)
E_{10}	6	8	(0)	7	7	(0)	11	3	(1)	11	3	(1)	8	6	(0)	11	3	(1)	11	3	(0)	9	5	(0)	74	38	(3)
Sub-Total	37	33	(3)	39	31	(1)	51	19	(4)	50	20	(4)	31	39	(1)	50	20	(1)	54	16	(0)	44	26	(0)	356	204	(14)
TOTALS:	70	70	(3)	73	67	(1)	110	30	(4)	106	34	(4)	53	87	(1)	93	47	(1)	112	28	(0)	90	50	(1)	707	413	(15)

NOTE: The first number in each column is the number of correct judgments made in that category; the second number is the number of incorrect judgments; the number in parentheses is the number of inconclusive judgments. The ordering of individual evaluators in this table is consistent with other tables in the appendices listing individual results.

TABLE D. 2.--Confidence Scores of Individual Evaluators.*

Evaluators	Categories of Record Sets							
	Verified				Unverified			
	Truthful		Deceptive		Truthful		Deceptive	
	Crime Against A Person	Property Crime	Crime Against A Person	Property Crime	Crime Against A Person	Property Crime	Crime Against A Person	Property Crime
Low Experience								
E_1	51	57	63	60	53	54	59	54
E_2	56	59	63	60	51	49	49	52
E_3	58	51	62	56	50	51	56	58
E_4	41	41	54	53	50	40	50	45
E_5	39	43	43	48	46	46	52	44
$\bar{X}$	49.0	50.2	57.0	55.4	50.0	48.0	53.2	50.6
High Experience								
E_6	53	52	55	52	51	49	54	52
E_7	57	55	60	60	59	55	61	58
E_8	63	64	66	71	62	68	72	63
E_9	41	49	47	52	48	57	52	55
E_{10}	53	55	57	62	54	54	58	55
$\bar{X}$	53.4	55.0	57.0	59.4	54.8	56.6	59.4	56.6
Combined $\bar{X}$	51.2	52.6	57.0	57.4	52.4	52.3	56.3	53.6

*The possible range for confidence scores per category = 14-84.

TABLE D.3.—Total Ease of Interpretability Scores of Individual Evaluators.*

	Category of Record Sets							
	Verified				Unverified			
	Truthful		Deceptive		Truthful		Deceptive	
Evaluators	Crime Against A Person	Property Crime	Crime Against A person	Property Crime	Crime Against A Person	Property Crime	Crime Against A Person	Property Crime
Low Experience								
E_1	116	123	144	134	121	123	130	122
E_2	107	110	125	111	117	98	103	101
E_3	109	98	129	113	117	102	119	115
E_4	87	90	112	103	107	84	109	97
E_5	85	82	87	96	79	84	89	85
$\bar{X}$	100.8	100.6	119.4	111.4	108.2	98.2	110.0	104
High Experience								
E_6	130	122	141	136	124	124	136	134
E_7	98	92	120	115	106	94	113	100
E_8	123	122	141	141	139	136	137	117
E_9	105	112	123	120	110	114	109	113
E_{10}	99	105	102	106	97	86	104	102
$\bar{X}$	111.0	110.6	125.4	123.6	115.2	110.8	119.8	113.2
Combined $\bar{X}$	105.9	105.6	122.4	117.5	111.7	104.5	114.9	108.6

*The possible range for total Ease-of-Interpretability scores per category = 42-210.

APPENDIX E

ANALYSIS OF VARIANCE TABLES

TABLE E.1.--Analysis of Variance Table for Accuracy Scores.

Source	df	MS	F	p<
1. A (A=Experience-high,low)	1	15.75	.11	.75
2. E (E=Evaluators): A	8	149.47		
3. B (B=Verification-verified, unverified)	1	77.42	1.42	.27
4. C (C=Truthfulness-truthful, deceptive)	1	10628.36	10.70	.01
5. D (D=Crime type-person, property)	1	183.32	1.54	.25
6. A X B	1	143.92	2.64	.14
7. A X C	1	1292.03	1.30	.29
8. A X D	1	5.57	.05	.83
9. B X C	1	183.92	1.36	.28
10. B X D	1	231.54	1.18	.31
11. C X D	1	3039.35	55.83	.0001
12. A X B X C	1	.63	.005	.95
13. A X B X D	1	5.78	.03	.87
14. A X C X D	1	15.93	.29	.60
15. B X C X D	1	1925.70	20.87	.002
16. A X B X C X D	1	5.57	.06	.81
17. B X E:A	8	54.46		
18. C X E:A	8	992.65		
19. D X E:A	8	119.05		
20. B X C X E:A	8	135.30		
21. B X D X E:A	8	195.76		
22. C X D X E:A	8	54.44		
23. B X C X D X E:A	8	92.29		

TABLE E.2.--Analysis of Variance Table for Confidence Scores.

Source	df	MS	F	p<
1. A (A=Experience-high,low)	1	470.45	1.77	.22
2. E (E=Evaluators):A	8	266.50		
3. B (B=Verification-verified, unverified)	1	16.20	.53	.49
4. C (C=Truthfulness-truthful, deceptive)	1	312.05	64.17	.0001
5. D (D=Crime type- person, property)	1	1.25	.08	.78
6. A X B	1	48.05	1.57	.25
7. A X C	1	12.80	2.63	.14
8. A X D	1	20.00	1.30	.29
9. B X C	1	36.45	6.23	.03
10. B X D	1	26.45	2.16	.18
11. C X D	1	16.20	2.72	.13
12. A X B X C	1	5.00	.85	.38
13. A X B X D	1	.20	.02	.90
14. A X C X D	1	.05	.008	.93
15. B X C X D	1	3.2	.42	.54
16. A X B X C X D	1	18.05	2.37	.16
17. B X E:A	8	30.69		
18. C X E:A	8	4.86		
19. D X E:A	8	15.38		
20. B X C X E:A	8	5.85		
21. B X D X E:A	8	12.26		
22. C X D X E:A	8	5.94		
23. B X C X D X E:A	8	7.63		

TABLE E.3.--Analysis of Variance Table for Total Ease of Interpretability Scores.

Source	df	MS	F	p<
1. A (A=Experience- high,low)	1	1852.81	1.03	.34
2. E (E=Evaluators):A	8	1800.91		
3. B (B=Verification-verified, unverified)	1	171.11	7.65	.02
4. C (C=Truthfulness- truthful, deceptive)	1	1593.11	37.99	.0003
5. D (D=Crime type- person property)	1	437.11	8.22	.02
6. A X B	1	.012	.0006	.98
7. A X C	1	2.11	.05	.83
8. A X D	1	37.81	.71	.42
9. B X C	1	556.51	9.31	.02
10. B X D	1	86.11	3.25	.11
11. C X D	1	17.11	1.05	.34
12. A X B X C	1	.61	.010	.92
13. A X B X D	1	.31	.012	.92
14. A X C X D	1	.013	.0008	.98
15. B X C X D	1	37.81	1.05	.34
16. A X B X C X D	1	49.61	1.38	.27
17. B X E:A	8	22.38		
18. C X E:A	8	41.93		
19. D X E:A	8	53.15		
20. B X C X E:A	8	60.94		
21. B X D X E:A	8	26.46		
22. C X D X E:A	8	16.31		
23. B X C X D X E:A	8	36.03		

TABLE E.4.--Analysis of Variance Table for Respiration Ease-of-Interpretability Scores.

Source	df	MS	F	p<
1. A (A=Experience-high, low)	1	11.25	.06	.81
2. E (E=Evaluators): A	8	187.94		
3. B (B=Verification-verified, unverified)	1	115.20	40.87	.0003
4. C (C=Truthfulness-truthful, deceptive)	1	510.05	102.65	.0001
5. D (D=Crime type-person, property)	1	31.25	4.17	.08
6. A X B	1	20.00	7.10	.03
7. A X C	1	2.45	.49	.50
8. A X D	1	6.05	.81	.40
9. B X C	1	72.20	6.25	.04
10. B X D	1	88.20	20.54	.002
11. C X D	1	36.45	5.33	.05
12. A X B X C	1	.20	.02	.90
13. A X B X D	1	.20	.05	.83
14. A X C X D	1	.05	.01	.93
15. B X C X D	1	.80	.07	.80
16. A X B X C X D	1	51.20	4.51	.07
17. B X E:A	8	2.82		
18. C X E:A	8	4.97		
19. D X E:A	8	7.49		
20. B X C X E:A	8	11.54		
21. B X D X E:A	8	4.29		
22. C X D X E:A	8	6.84		
23. B X C X D X E:A	8	11.34		

TABLE E.5.--Analysis of Variance Table for GSR Ease-of-Interpretability Scores.

Source	df	MS	F	p<
1. A (A=Experience-high, low)	1	312.05	.78	.40
2. E (E=Evaluators) : A	8	401.00		
3. B (B=Verification-verified, unverified)	1	14.45	2.63	.14
4. C (C=Truthfulness-truthful, deceptive)	1	5.00	.59	.46
5. D (D=Crime type-person, property)	1	22.05	3.82	.09
6. A X B	1	1.80	.33	.58
7. A X C	1	.45	.05	.82
8. A X D	1	5.00	.87	.38
9. B X C	1	140.45	12.13	.008
10. B X D	1	12.80	3.07	.12
11. C X D	1	11.25	4.79	.06
12. A X B X C	1	3.20	.28	.61
13. A X B X D	1	6.05	1.45	.26
14. A X C X D	1	3.20	1.36	.28
15. B X C X D	1	45.00	6.29	.04
16. A X B X C X D	1	6.05	.85	.38
17. B X E:A	8	5.50		
18. C X E:A	8	8.48		
19. D X E:A	8	5.78		
20. B X C X E:A	8	11.58		
21. B X D X E:A	8	4.18		
22. C X D X E:A	8	2.35		
23. B X C X D X E:A	8	7.15		

TABLE E.6.--Analysis of Variance Table for Cardio Ease-of-Interpretability Scores.

Source	df	MS	F	p<
1. A (A=Experience-high, low)	1	485.11	3.37	.10
2. E (Evaluators) :A	8	143.91		
3. B (B=Verification-verified, unverified)	1	2.11	.17	.69
4. C (C=Truthfulness-truthful, deceptive)	1	227.81	59.27	.0001
5. D (D=Crime type-person, property)	1	112.81	9.87	.01
6. A X B	1	9.11	.72	.42
7. A X C	1	.31	.08	.78
8. A X D	1	2.11	.18	.68
9. B X C	1	10.51	1.48	.26
10. B X D	1	12.01	2.12	.18
11. C X D	1	2.11	.37	.56
12. A X B X C	1	4.51	.64	.45
13. A X B X D	1	12.01	2.12	.18
14. A X C X D	1	2.81	.49	.50
15. B X C X D	1	.11	.02	.89
16. A X B X C X D	1	5.51	.93	.36
17. B X E:A	8	12.64		
18. C X E:A	8	3.84		
19. D X E:A	8	11.43		
20. B X C X E:A	8	7.11		
21. B X D X E:A	8	5.67		
22. C X D X E:A	8	5.74		
23. B X C X D X E:A	8	5.91		

APPENDIX F

CORRELATIONS OF EVALUATORS' NUMERICAL SCORES ON VERIFIED AND UNVERIFIED RECORD SETS

TABLE F.1.--Correlations of Respiration Scores: Verified Record Sets.

Evaluator	1	2	3	4	5	6	7
1		.68	.24	.76	.52	.79	.62
2			.15	.65	.43	.76	.61
3				.06	.55	.41	.52
4					.55	.77	.56
5						.59	.89
6							.74

TABLE F.2.--Correlations of Respiration Scores: Unverified Record Sets.

Evaluator	1	2	3	4	5	6	7
1		.65	.58	.66	.44	.43	.73
2			.49	.75	.49	.62	.66
3				.83	.62	.35	.89
4					.69	.58	.91
5						.41	.77
6							.53

TABLE F.3.--Correlations of GSR Scores: Verified Record Sets.

Evaluator	1	2	3	4	5	6	7
1		.87	.63	.75	.86	.88	.73
2			.67	.68	.74	.88	.54
3				.68	.66	.70	.58
4					.71	.76	.64
5						.75	.64
6							.60

TABLE F.4.--Correlations of GSR Scores: Unverified Record Sets.

Evaluator	1	2	3	4	5	6	7
1		.74	.70	.87	.76	.81	.82
2			.66	.76	.59	.77	.76
3				.75	.64	.63	.63
4					.81	.85	.75
5						.77	.64
6							.69

TABLE F.5.--Correlations of Cardio Scores: Verified Record Sets.

Evaluator	1	2	3	4	5	6	7
1		.67	.41	.66	.59	.37	.70
2			.48	.71	.62	.54	.74
3				.32	.30	.41	.62
4					.80	.58	.79
5						.51	.71
6							.57

TABLE F.6.--Correlations of Cardio Scores: Unverified Record Sets.

Evaluator	1	2	3	4	5	6	7
1		.61	.61	.63	.70	.44	.79
2			.60	.76	.67	.77	.54
3				.35	.63	.43	.52
4					.77	.59	.70
5						.69	.88
6							.53

TABLE F.7.--Correlations of Combined Scores: Verified Record Sets.

Evaluator	1	2	3	4	5	6	7
1		.77	.35	.71	.65	.75	.68
2			.43	.69	.56	.79	.63
3				.26	.54	.48	.57
4					.70	.72	.71
5						.57	.85
6							.67

TABLE F.8.--Correlations of Combined Scores: Unverified Record Sets.

Evaluator	1	2	3	4	5	6	7
1		.66	.72	.73	.63	.48	.79
2			.62	.84	.65	.70	.66
3				.70	.69	.47	.83
4					.78	.65	.83
5						.57	.83
6							.50

BIBLIOGRAPHY

BIBLIOGRAPHY

Books

Arther, R. and Caputo, R., Interrogation for Investigators. New York: William C. Copp and Associates, 1959.

Chao, L., Statistics: Methods and Analyses. New York: McGraw-Hill, 1969.

Ferguson, R., The Scientific Informer. Springfield, Illinois: Charles C Thomas, 1971.

Greenfield, N. and Sternbach, R. (eds.), Handbook of Psychophysiology. New York: Holt, Rinehart and Winston, 1972.

Harrelson, L., Keeler Polygraph Institute Training Guide. Chicago: Keeler Polygraph Institute, 1964.

Inbau, F. and Reid, J., Lie Detection and Criminal Interrogation. Baltimore, Maryland: Williams and Wilkins, 1953.

Kirk, R., Experimental Design: Procedures for the Behavioral Sciences. Belmont, California: Brooks/Cole, 1968.

Larson, J., Lying and Its Detection. Chicago: University of Chicago Press, 1932. Reprinted, Montclair, New Jersey: Patterson Smith, 1969.

Lee, C., The Instrumental Detection of Deception: The Lie Test. Springfield, Illinois: Charles C Thomas, 1953.

Leonard, V.A. (ed.), Academy Lectures on Lie Detection. Springfield, Illinois: Charles C Thomas, 1957.

Leonard, V.A. (ed.), Academy Lectures on Lie Detection. Vol. 2, Springfield, Illinois: Charles C Thomas, 1958.

Lykken, D., Psychology and The Lie Detector Industry. Minneapolis: Department of Psychiatry, University of Minnesota Press, Report No. PR-74-1, 1974.

Marston, W., The Lie Detector Test. New York: Richard K. Smith, 1938.

Mehrens, W.A. and Ebel, R. (eds.), Principles of Educational and Psychological Measurement. Chicago: Rand McNally, 1967.

Munsterberg, H., On The Witness Stand. New York: Doubleday, 1908.

Prokasy, W.F. and Raskin, D. (eds.), Electrodermal Activity in Psychological Research. New York: Academic Press, in press.

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

Rosenthal, R., Experimenter Effects in Behavioral Research. New York: Appleton-Century-Crofts, 1966.

Periodicals

Abrams, S., "Polygraph Validity and Reliability: A Review," Journal of Forensic Sciences, 18 (1973), 313-326.

Alpert, M., Kurtzberg, R.L. and Friedhoff, A., "Transient Voice Changes Associated With Emotional Stimuli," Archives of General Psychiatry, 8 (1963), 362-365.

Altarescu, H., "Problems Remaining for the 'Generally Accepted' Polygraph," Boston University Law Review, 53 (1973), 375-405.

Ansley, N. (ed.), "Actions of the Board of Directors, January 18-20," American Polygraph Association Newsletter, 1 (1974), 10.

Ansley, N. (ed.), "A.P.A. Accepted Polygraph Schools," American Polygraph Association Newsletter (December/ January, 1974), 14.

Ansley, N. (ed.), "Inquiry Regarding Dektor PSE-1," American Polygraph Association Newsletter, 3 (1972), 18.

Arther, R., "Covering Two Crimes in One Examination," Journal of Polygraph Studies, 4 (1970), 3-4.

Arther, R., "Crime Question Wording," Journal of Polygraph Studies, 4 (1969), 1-4.

Arther, R., "Irrelevant Questions," Journal of Polygraph Studies, 3 (1969), 3-4.

Arther, R., "Peak of Tension: Basic Information," Journal of Polygraph Studies, 1 (1967), 4.

Arther, R., "Peak of Tension: Dangers," Journal of Polygraph Studies, 2, 5 (1968), 1-4.

Arther, R., "Peak of Tension: Examination Procedures," Journal of Polygraph Studies, 5, 1 (1970), 1-4.

Arther, R., "The Guilt Complex Question," Journal of Polygraph Studies, 4 (1969), 1-4.

Backster, C., "Methods of Strengthening Our Polygraph Technique," Police, 6, 5 (1962), 61-68.

Backster, C., "Lie Detection Comes of Age," Law and Order (undated, unpaginated reprint supplied by author).

Ben Shakhar, G., Lieblich, I., and Kugelmass, S., "Guilty Knowledge Technique: Application of Signal Detection Measures," Journal of Applied Psychology, 54, 5 (1970), 409-413.

Benussi, V., "On the Effects of Lying on Changes in Respiration," Archives Fur Die Gesamte Psychologie (1914), 244-273.

Berkhout, J., Walter, D., and Adey, W., "Autonomic Responses During A Replicable Interrogation," Journal of Applied Psychology, 54, 4 (1970), 316-325.

Berrien, F., "Possibilities in the Use of the Ophthalmograph as a Supplement to Existing Indices of Deception," Psychological Bulletin (Abstract), 37 (1940), 507.

Berrien, F., "Ocular Stability in Deception," Journal of Applied Psychology, 26 (1942), 55-63.

Berrien, F., and Huntington, G., "An Exploratory Study of Pupillary Responses During Deception," Journal of Experimental Psychology, 32 (1943), 443-449.

Bersh, P., "A Validation Study of Polygraph Examiner Judgments," Journal of Applied Psychology, 53, 5 (1969), 399-403.

Bitterman, M., and Marcuse, F., "Cardiovascular Responses of Innocent Persons to Criminal Interrogation," American Journal of Psychology, 60 (1947), 407-412.

Brisentine, R., "Quality Control," Polygraph, 2 (1973, 278-286.

Burtt, H., "Further Technique for Inspiration Expiration Ratios," Journal of Experimental Psychology, 4 (1921), 106-110.

Burtt, H., "The Inspiration-Expiration Ratio During Truth and Falsehood," Journal of Experimental Psychology, 4, 1 (1921), 1-23.

Chappell, N., Matthew, N., "Blood Pressure Changes in Deception," Archives of Psychology, 17, 105 (1929), 1-39.

Davidson, P., "Validity of the Guilty-Knowledge Technique: The Effects of Motivation," Journal of Applied Psychology, 52, 1 (1968), 62-65.

Dearman, H., and Smith, B., "Unconscious Motivation and the Polygraph Test," American Journal of Psychiatry, 119, 11 (1963), 1017-1021.

Fay, P., and Middleton, W., "The Ability to Judge Truth-Telling or Lying From the Voice as Transmitted Over a Public Address System," Journal of General Psychology, 24 (1941), 211-215.

Geldreich, E., "Studies of the Galvanic Skin Response As a Deception Indicator," Transactions Kansas Academy of Sciences, 44 (1941), 346-351.

Gustafson, L., and Orne, M., "Effects of Heightened Motivation on the Detection of Deception," Journal of Applied Psychology, 47, 6 (1963), 408-411.

Gustafson, L., and Orne, M., "The Effects of Task and Method of Stimulus Presentation on the Detection of Deception," Journal of Applied Psychology, 48, 6 (1964), 383-387.

Gustafson, L., and Orne, M., "The Effects of Verbal Responses on the Laboratory Detection of Deception," Psychophysiology, 2, 1 (1965), 10-13.

Harmon, G., and Reid, J., "The Selection and Phrasing of Lie-Detector Test Control - Questions," Journal of Criminal Law, Criminology and Police Science, 46 (1955), 578-582.

Heckel, R., Brokaw, J., Salzburg, H., and Wiggins, S., "Polygraphic Variations in Reactivity Between Delusional, Non-Delusional and Control Groups in a 'Crime' Situation," Journal of Criminal Law, Criminology and Police Science, 53, 3 (1962), 380-383.

Horvath, F., "Verbal and Nonverbal Clues to Truth and Deception During Polygraph Examinations," Journal of Police Science and Administration, 1, 2 (1973), 138-152.

Horvath, F., and Reid, J., "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," Journal of Criminal Law, Criminology and Police Science, 62, 2 (1971), 276-281.

Horvath, F., and Reid, J., "The Polygraph Silent Answer Test," Journal of Criminal Law, Criminology and Police Science, 63, 2 (1972), 285-293.

Hunter, F., and Ash, P., "The Accuracy and Consistency of Polygraph Examiner's Diagnoses," Journal of Police Science and Administration, 1 (1973), 370-375.

Keeler, L., "A Method for Detecting Deception," The American Journal of Police Science, 1 (1930), 38-52.

Kubis, J., "Electronic Detection of Deception," Electronics, 18 (1945), 192-212.

Kubis, J., "Experimental and Statistical Factors in the Diagnosis of Consciously Suppressed Affective Experiences," Journal of Clinical Psychology, 6 (1950), 12-16.

Kugelmass, S., and Lieblich, I., "Effects of Realistic Stress and Procedural Interference in Experimental Lie Detection," Journal of Applied Psychology, 50, 3 (1966), 211-216.

Kugelmass, S., Lieblich, I., and Bergman, Z., "The Role of Lying in Psychophysiological Detection," Psychophysiology, 3, 3 (1967), 312-315.

Kugelmass, S., Lieblich, I., Ben-Ishai, A., Opatowski, A., and Kaplan, M., "Experimental Evaluation of Galvanic Skin Response and Blood Pressure Change Indices During Criminal Interrogation," Journal of Criminal Law, Criminology and Police Science, 59, 4 (1968), 632-635.

Landis, C., "Electrical Phenomenon of the Skin," Psychological Bulletin, 29, 10 (1932), 693-752.

Landis, C., and Gullette, R., "Studies of Emotional Reactions," Journal of Comparative Psychology, 5 (1925), 221-253.

Landis, C., and DeWick, H., "The Electrical Phenomenon of the Skin (Psychogalvanic Reflex)," Psychological Bulletin, 26, 1 (1929), 64-119.

Landis, C., and Wiley, L., "Changes of Blood Pressure and Respiration During Deception," Journal of Comparative Psychology, 6 (1926), 1-19.

Larson, J., "Modification of the Marston Deception Test," Journal of the American Institute of Criminal Law and Criminology, 12 (1921), 390-399.

Larson, J., "The Cardio Pneumo Psychogram and Its Use in the Study of Emotions, with Practical Applications," Journal of Experimental Psychology, 5 (1922), 323-328.

Lykken, D., "The GSR in the Detection of Guilt," Journal of Applied Psychology, 43, 6 (1959), 385-388.

Lykken, D., "The Validity of the Guilty Knowledge Technique: The Effects of Faking," Journal of Applied Psychology, 44, 4 (1960), 258-262.

Lyon, V., "Deception Tests with Juvenile Delinquents," Journal of General Psychology, 48 (1936), 494-497.

MacNitt, R., "In Defense of the Electrodermal Response and Cardiac Amplitude as Measures of Deception," Journal of Criminal Law and Criminology, 33, 3 (1942), 266-275.

Marston, W., "Systolic Blood Pressure Symptoms and Deception," Journal of Experimental Psychology, 2 (1917), 117-163.

Marston, W., "Psychological Possibilities in the Deception Test," Journal of The American Institute of Criminal Law and Criminology, 2, 4 (1921), 551-570.

Obermann, C., "The Effect on the Berger Rhythm of Mild Affective States," Journal of Abnormal and Social Psychology, 34 (1939), 84-95.

Orne, M., "Implications of Laboratory Research for the Detection of Deception," Polygraph, 2 (1973), 169-199.

Paterson, R., "The Future of Polygraph in Industrial Security," American Polygraph Association Newsletter, No. 8 (1972), 1-3.

Peterson, F., and Jung, C., "Psychophysical Investigations with the Galvanometer and Pneumograph in Normal and Insane Individuals," Brain, 30 (1907), 153-218.

Reid, J., "Simulated Blood Pressure Responses in Lie-Detector Tests and a Method for Their Detection," Journal of Criminal Law and Criminology, 36, 1 (1945), 201-214.

Reid, J., "A Revised Questioning Technique in Lie-Detector Tests," Journal of Criminal Law and Criminology and American Journal of Police Science, 37, 6 (1947), 542-547.

Reid, J., and Arther, R., "Behavior Symptoms of Lie Detector Subjects," Journal of Criminal Law, Criminology and Police Science, 44, 1 (1953), 104-108.

Romig, C., "The Status of Polygraph Legislation of the Fifty States," Police, 16, 2 (1971), 54-61.

Ruckmick, C., "The Truth About the Lie Detector," Journal of Applied Psychology, 22, 1 (1938), 50-58.

Sternbach, R., Gustafson, L., and Colier, R., "Don't Trust the Lie Detector," Harvard Business Review, 40, 6 (1962), 127-134.

Summers, W., "Science Can Get The Confession," Fordham Law Review, 8 (1939), 334-354.

Suzuki, A., "An Analysis of Relative Effectiness (sic) of the Physical Indices and the Influence of Polygraph Examiner's Experience Upon Judgment of Polygraph Records in Detection of Deception," Japanese Journal, (title unknown), 21, 3 (1968), 51-59.

Thackray, R., and Orne, M., "A Comparison of Physiological Indices in Detection of Deception," Psychophysiology, 4, 3 (1968), 329-339.

Thackray, R., and Orne, M., "Effects of the Type of Stimulus Employed and the Level of Subject Awareness on the Detection of Deception," Journal of Applied Psychology, 52, 3 (1968), 234-239.

Trovillo, P., "Deception Test Criteria," Journal of Criminal Law and Criminology, 33 (1942), 338-358.

Trovillo, P., "A History of Lie Detection," Journal of Criminal Law, Criminology and Police Science, 29 (1939), 848-881 and 30, 104-119.

Van Buskirk, D., and Marcuse, F., "The Nature of Errors in Experimental Lie Detection," Journal of Experimental Psychology, 47 (1954), 187-190.

Unpublished Works

Barland, G., An Experimental Study of Field Techniques in Lie Detection (unpublished Master's Thesis, Department of Psychology, University of Utah, 1972).

Moroney, W., The Detection of Deception as a Function of PGR Methodology (unpublished Ph.D. dissertation, St. John's University, 1968. Ann Arbor, Michigan: University Microfilms, 1969, No. 69-7125).

Reid, J., Interpretation of Truth and Deception in Polygraph Test Records (Undated, unpublished manuscript supplied by author).

Reid, J., Stimulation Technique Outline, undated, unpublished manuscript supplied by J.E. Reid and Associates, Chicago.

Rouke, F., Evaluation of the Indices of Deception in the Psychogalvanic Technique (unpublished Ph.D. dissertation, Fordham University, 1941).

Scheifley, Verda, and Schmidt, W., Jeremy D. Finn's Multivariance-Univariate and Multivariate Analysis of Variance, Covariance and Regression, occasional paper No. 22, Office of Research Consultation, Michigan State University, 1973.

Government Documents

Ellson, D., Davis, R., Burke, C., and Saltzman, I., A Report of Research on Detection of Deception, prepared for Office of Naval Research, Contract N60Nr-18011, Department of Psychology, University of Indiana, 1952.

Federal Bureau of Investigation, Uniform Crime Reports for the United States: 1972, Washington: Government Printing Office, 1973.

Kubis, J., Studies in Lie Detection: Computer Feasibility Considerations, Technical Report 62-205, Arlington, Virginia: Armed Services Technical Information Agency, 1962, prepared for Air Force Systems Command, Contract No. AFBO (602)-22700, Project No. 5534, Fordham University, 1962.

Kugelmass, S., Effects of Three Levels of Realistic Stress on Differential Psychological Reactivities, AFEOAR Grant 63-61, Air Force Office of Scientific Research, European Office, Aerospace Research, U.S. Air Force, Hebrew University of Jersusalem, Isreal, 1963.

Orlansky, J., An Assessment of Lie Detection Capability (Declassified Version), Technical Report 62-16, Arlington, Virginia: Institute for Defense Analyses, Research and Engineering Support Division, 1964.

U.S. Congress, House, Subcommittee of the Committee on Government Operations, Use of Polygraphs as "Lie Detectors" by the Federal Government, Hearings, 88th Congress, 2nd Session, and 89th Congress, 1st Session, Parts 1-6, Washington, D.C.: U.S. Government Printing Office, 1964-1966.

Violante, R., and Ross, S., Research on Interrogation Procedures, Interim Report prepared for U.S. Navy, Office of Naval Research, Contract Nonr 4129(00), Stanford Research Institute, Menlo Park, California, 1964.

Other Sources

Arther, R., "The Heart and You" (unpublished, undated manuscript, National Training Center of Lie Detection, New York).

Backster, C., Standardized Polygraph Notepack and Technique Guide, New York: Backster Research Foundation, 1969.

Backster, C., Tri-Zone Polygraph, New York: Backster Research Foundation, 1969.

Barland, G., The Reliability of Polygraph Chart Evaluation, paper presented to American Polygraph Association Seminar in Chicago, Illinois, 1972.

Golden, R., The "Yes"-"No" Technique, paper presented to the American Polygraph Association Annual Convention in Houston, Texas, 1969.

Klump, C., Principles of Controlled Stimulation, paper presented at American Academy of Polygraph Examiners, Eighth Annual Seminar, Washington, D.C., 1961.

Orne, M., Untitled Manuscript, presented to American Polygraph Association, Third Annual Seminar, Silver Springs, Maryland, 1969.