Chapter 5

Review and Analysis of Polygraph Analog Studies

Chapter 5

Review and Analysis of Polygraph Analog Studies

INTRODUCTION

Analog studies, for purposes of the present analysis, are investigations in which field methods of polygraph examinations are used in simulated criminal or other situations. Such studies investigate either "mock" crimes set up by an experimenter (with the knowledge and collaboration of subjects) or actual small crimes "induced" by the experimenter. Such analog studies are not actual criminal investigations and subjects are usually aware that they are participants in polygraph research. Analog studies differ from other laboratory studies of polygraph testing in that they simulate actual field examinations. However, in analog studies, typical components of field examinations are replicated to the extent it is possible to do so. Such studies test the validity of various polygraph techniques under controlled conditions. In chapter 4, the results of a systematic review of field studies of validity were presented. In the present chapter, a similar analysis of analog studies is presented. As with the field studies, the studies concern the use of polygraph examinations for investigation of crimes. The two exceptions (16,43) use analogs to the type of relevant/irrelevant (R/1) question technique typically used in the personnel screening situation.

The present chapter is organized as follows: first, the characteristics of analog studies and the varieties of ways in which they differ from field studies are discussed. Then, the criteria used for including studies in the analysis are described. The coding procedure, which is essentially the same as that used to code the field studies, is described briefly. Analog studies of the control question technique (CQT), guilty knowledge technique (GKT), and personnel screening examination are then reviewed. The findings of a statistical analysis of the analog studies complete the chapter.

CHARACTERISTICS OF ANALOG STUDIES

The "crimes" utilized in analog studies in order to establish ground truth have taken different forms. For the most part, they are "mock crimes;" i.e., crimes in which subjects know they are "role playing" at being criminals for purposes of an experiment. Mock crime studies may be further differentiated by whether or not the experimenter controls the guilt or innocence of research participants. In some studies, subjects know that the crime is part of the experimental situation but they are more or less free to go through with the crime or not. Two analog studies have utilized actual small crimes. In these studies, apparently real situations were embedded in an experimental situation in which subjects were given an opportunity to commit a crime or not.

The consequences of failing a polygraph examination (e. g., a possible prison sentence) cannot be replicated in the laboratory. In analog studies, punishment takes such forms as losing the chance for a monetary reward. Some researchers have experimented with other punishments such as electric shock *(105)* or the threat of shocks (35). The analog studies that use real crimes provide another alternative, in that subjects can be threatened with real punishment (e. g., academic sanctions for cheating on an examination). In still other cases, subjects are led to believe that "stable" individuals can avoid detection.

Analog studies represent, thus, a "tradeoff" to the investigator interested in polygraph testing

validity. On the one hand, because the researcher sets up the crime, ground truth is known; and because "ground truth" is established, analog studies are superior to field studies in terms of criterion validity. Furthermore, they provide the investigator with more control of the polygraph situation and conditions of testing. The experimenter can select particular subject groups, can standardize testing procedures for all subjects, and can systematically vary guilt or innocence. With this control, the experimenter can also directly compare the effects of variations in polygraph techniques, physiological measures, information given to subjects, and scoring methods.

On the other hand, although analog studies have greater criterion validity and offer greater experimental control, their use as indicators of polygraph testing validity is potentially problematic. The reasons have to do primarily with external validity (20,136; see, also, 1,7,108); i.e., the crime situation differs, the testing situations in the field and the laboratory differ, the training of the examiners differs, the subject population differs, and, apparently most important, the consequences for "suspects" differ dramatically between the field and the laboratory. In addition, in analog studies, the questions and question techniques most often are not tailored to individual subjects. In actual criminal field investigations, case information about the crime and the subject usually provides a basis for tailoring questions.

Numerous specific differences can be noted. Perhaps most importantly, the laboratory crime and the consequences of detection are much less serious. In addition, in an analog study, demand characteristics (which suggest to the subject desirable responses) may create a somewhat different polygraph situation than found in typical field situations *(20).* In terms of factors that may increase validity of analog studies, there is some evidence that laboratory researchers are, in general, able to use more sophisticated and stable equipment than portable machines often used in the field (136). On the other hand, examinations in analog studies are often conducted by researchers who are primarily psychophysiologists (e. g., 49) or psychologists (43) with only limited training in field techniques. Field examinations, in contrast, are conducted by individuals whose primary training is as polygraph examiners and who are usually experienced. This would suggest that field examinations may be more accurate.

The characteristics of subjects who participate in analog studies also vary from subjects in field studies. Several use college students, others recruit community members through the newspaper, one uses police candidates, and another prison inmates. In many studies, subjects are probably better educated and more highly socialized than the average field examinee. In the case of student subjects, they are probably younger on the average and from a higher social class as well. Raskin (132) notes that analog studies using students yield a lower accuracy rate than other studies. As will be discussed below, this may be due to subject differences between field and analog studies because a realistic fear of failure does not play a central role for subjects. The consequences of failure for analog studies are usually minimal in contrast to typical field investigations.

Study Selection

For present purposes, studies were only included as analog for the primary analyses if they employed actual field polygraph techniques to detect deception or concealed information, and if the studies pertained to some use of polygraph testing in the real world. The studies selected are listed in tables *6* and 7. Studies of components of the polygraph examinations, such as studies which used only card tests *(97,101),* number tests *(120),* or tests concerning concealed personal information (e.g., parents' first name; see, e.g., 106) were not included.

In addition, studies were excluded because their primary focus was on a theoretical factor thought to affect validity, such as variability in physiological recordings (45), nonstandard means of interpreting such recordings (163), or the role of "lying" (96). Such studies will be referred to as laboratory investigations and are distinguished from analog studies.

Analog studies of the guilty knowledge test (GKT) have been included, although analyzed separately, because this form of the polygraph examination represents an alternative proposed for

Table 6.–Outcomes of Control Question Analog Studies of Validity

	Guilty				Innocent					
	Number of subjects	Correct	Incorrect (False negative)	Inconclusive	Nu-mber of subjects	Correct	Incorrect (False negative)	Inconclusive	Total number of subjects	Lambda[a]
Barland and Raskln (21)	36	63.90/.	8.30/.	27.80/o	36	41.70/0	16.70/0	41 .7%	72	047
Podlesny and Raskln (127)	20	69.0	16.0	15.0	20	910	4.0	5.0	.40	07'5
Raskln and Hare (137)	24	875	0	12.5	24	75.0	4.0	20.8	48	0.83
Rovner, et ala (143)	36	77.8	8.3	13.9	36	80.5	13.9	5.5	72	0.72
Klrcher (89a)	50	60.0	4,0	36.0	50	76.0	2.0	22.0	100	0.72
Dawson (49)	12	91.7	0	8.3	12	583	25.0	167	24	0.33
Widackl and Homath[b,e] (207)	20	90.0	50	5.0	—	—	—	—	20	—[k]
Bradley and Janisse (35)	96	—	—	—	96	—	—	—	192	0.33
EDR		60.4	13.5	26.0	—	58.3	9.4	32.3		
Heart rate		35.4	20.8	438	—	33.3	19.8	46.9		
Szucko and Kleinmuntz[c] (160)	15	71.3	28.7	[d]	15	49.3	50.7	[d]	30	022
Glnton, et al. (62)	2	100.0	0	C	13	84.6	15.4	0	15	0.00J
Honts and Hodes (75)[f]:										
No countermeasures	9	67.0	0.0	33.0	12	33.0	17.0	500		0.25
Countermeasures	[g]f	58.0	5.5	36.6	—9	—	—	—		—
	28				12				40	
Honts and Hodes (76):										
No countermeasures	19	84.2	0.0	15.8	19	31.6	15.8	52.6		0.53
Countermeasures	19	36.8	26.3	36.8	— 9	—	—	—		—
	38				19				57	
Heckel, et al. (74)										
Normals	—[h]				5	100.0	0.0	0.0		— k
Nondelusional psychiatric	— h				5	700	10.0	20.0		
Delusional psychiatric	- h				5	45.0	35.0	20.0	15	
Hammond (74a)	32	71.9	3.0	250	30	40.0	20.0	40,0	62	

a Summed across conditions
b **Examiner's task was to detect the one guilty person** in each of 20 **groups of four suspects.**
c **Based on ratings of 5* on a 1 to 8** scale of certainty of nondeception/deception
d Examiners were not allowed to categorize an examination as inconclusive
e Original subject assignments, 12 to each of 4 (including 2 countermeasure) conditions A total of five countermeasure subjects were eliminated from the analysis of results for guilty subjects for failure to follow countermeasure instructions Three no countermeasure subjects were elimi nated for spontaneously us l ng countermeasures
f Nine guilty subjects used pain countermeasures (tongue biting) and 10 used a muscle (toe pressing) countermeasure
g Innocent subjects used no countermeasures
h There was no guilty condition
i Not Included in analysis reported in table 8
j Lambda is a poor statistic when the base rate is skewed
k Lamba was not calculated when only guilty Or Innocent ***subjects*** were used

Table 7.—Outcomes of Guilty Knowledge Analog Studies of Validity

	Guilty				Innocent					
	Number of subjects	Correct	Incorrect (false negative)	Inconclusive	Number of subjects	Correct	Incorrect (false positive)	Inconclusive	Total number of subjects	Lambda[b]
Lykken[a] (105)	50	88.00/0	12.00/0	—	48	1OO.O°/;	0"/0	—	98	—[c]
Davidson (46)	12	91.7	8.3	—	36	100.0	0	—	48	0.92
Podlesny and Raskin (127)	10	80.0	**20.0**	**0**	10	80.0	0	**20.0**	**20**	**0.80**
Balloun and Holmes (12)	18			—					34	
Test 1		61.1	38.9	—		87.5	12.5	—		0.44
Test 2		16.7	83.3	—		93.7	6.3	—		
Giesen and Rollison (61)	20	95.0	5.0	—	20	100.0	0	—	40	0.95
Bradley and Janisse (35)	96			—	96				192	
EDR		59.4	40.6	—		88.5	11.5	—		0.38
Heart rate		44.8	55.2	—		82.3	17.7	—		
Timm (163)	237								237	
Liberal cutoff		80.8	19.2	—	–[b]	–	—	—		–[d]
Conservative cutoff		70.4	29.6	—	—	—	—	—		—

[a] Frequencies for detection of two mock crimes were combined,
[b] There were no innocent subjects
~Lambda cannot be calculated because crimes were not reported separately
Lambda cannot be calculated with only one condition,

use in the field (92,107,108), even though it has not been put into general practice.

Description of Studies

The following sections discuss each of the analog studies organized into three categories according to questioning technique. The discussion of CQT analog studies is first. Studies of CQT represent available studies, much like the case for field investigations (see ch. 4). Six studies of the concealed information or GKT and two of R/I follow. In only one study *(16),* involving the R/I technique, were subjects Government employees. The results of individual studies are summarized in tables 6 (CQT) and 7 (GKT). The description of the studies is followed by a systematic statistical analysis of the results of the CQT and GKT studies. The R/I studies were not analyzed as a group because of the paucity of studies.

Essentially, as shown in tables *6* to *9* the analysis of the analog studies yields conclusions similar to those of the field study analysis—i. e., although there is a greater-than-chance probability of detecting deceptive and nondeceptive subjects, there is what might be regarded as a significant error rate, and a great deal of variation across studies. However, as has been found in some reviews (1, 7), analog studies of CQT had lower accuracy rates than field studies of CQT.

In the studies detailed below, some experiments also tested the effect of factors hypothesized to have an effect on validity. For example, Barland and Raskin (22) examined the effect on validity of different types of feedback about the polygraph, and Dawson *(49)* investigated the effects of countermeasures. These factors are examined more systematically in chapter *6;* the emphasis of the present chapter is on the validity of different forms of polygraph examinations.

Table 8.—Mean Detection Rates as a Percentage of Total in Analog Studies of Control Question Technique

	Ground truth	
	Percent guilty	Percent innocent
Examiners' diagnosis	Mean	Mean
Deceptive	330	6.8
Nondeceptive	5.4	27.9
Inconclusive	13.4	13.5
	51.8	48.2

NOTE: $lambda_s$ = 0,43.

Table 9.—Mean Detection Rates as a Percentage of Total in Analog Studies of Guilty Knowledge Test

	Ground truth	
	Percent guilty	Percent innocent
Examiners' diaanosis	Mean	Mean
Guilty	27.9	2.2
Not guilty	17.3	52.6
Inconclusive	0	0
	45.2	54.8

NOTE: $lambda_s$ = O 70

CONTROL QUESTION TECHNIQUE

Fourteen analog studies of the control question technique were located. The largest group of these studies emanate from the research program of Professor David C. Raskin at the University of Utah. The remaining studies were conducted at a number of settings in the United States and elsewhere. Raskin and colleagues have conducted a systematic analog research program, and these studies are described as a group. Other researchers have published individual studies testing specific hypotheses relevant to the validity of the poly - graph. A description of these studies follows discussion of the University of Utah studies.

University of Utah Studies

Despite longstanding controversy about polygraph validity, the first research project conducting an analog study that simulated field polygraph techniques was not conducted until the 1970's (136). It was then that an ongoing research program headed by Professor Raskin at the Univer-

sity of Utah began to study the validity of the polygraph through analog experiments. In addition, these studies also examined the relationship to validity of different polygraph techniques (e.g., the stimulation test), different physiological measures, different methods of assessing the results, different types of information provided to subjects, and different subject and situation factors that could potentially affect polygraph validity.

The experiments conducted by Raskinandcolleagues use similar procedures to setup the mock crime and to conduct polygraph testing. In each of their studies, subjects are randomly assigned to an "innocent" condition or to a "guilty" condition. The mock crime is the theft of a small amount of money or a valuable object from a desk in a nearby room. To increase their motivation, subjects are offered a financial bonus for convincing the examiner they are innocent. In the testing the examiner employs the Federal zone of comparison (ZOC) control question technique, including a pretest interview. A numerical field scoring method developed by the Utah group (21) is used to make the diagnosis of truthfulness or deception.

Barland and Raskin

In the initial analog study using CQT (21), 72 student "guilty" and "innocent" volunteers were randomly assigned to one of three "feedback" conditions. The positive feedback subjects were instructed that the polygraph was effective, the negative feedback students were told that the machine was not working properly, and the other students received no feedback. Subjects then underwent a complete polygraph examination including a pretest interview. The Federal version of the ZOC technique was employed, with standard control questions used for all subjects. On average, the CQT identified 53 percent of all subjects correctly. Twelve percent were identified incorrectly and 35 percent of the examinations were inconclusive. Of the errors, three (4 percent of the entire sample) were false negatives and six (8 percent) were false positives.

Podlesny and Raskin

Podlesny and Raskin (127) conducted a more extensive experiment to examine the accuracy of CQT using three different types of control questions. They also tested the accuracy of behavioral observations of the examinee (80,139) in detecting deception, since this type of information is used in many field examinations and could possibly affect the validity of the technique (107,108). They compared as well the capability of different physiological measures in differentiating between guilty and innocent subjects. A GKT was also conducted with *20* subjects (see below).

In Podlesny and Raskin's study, subjects were community members who responded to newspaper advertisements. The experimenters drew from the Reid method in their design of the pretest interview (see ch. 2). One experimenter asked the subjects three questions from Reid and Horvath's structured pretest interview designed to provoke the subjects into displaying "behavioral symptoms" of deception *(80,139).*

During the polygraph examination they included two special types of control questions among the set of questions asked of the subjects. One was a "guilt complex question, " which asked the subject if he committed a fictitious crime of the same nature as the real crime. In this study, the guilt complex question was, "Did you take that watch from room *702?" (127).* There was, of course, no watch stolen from room *702.* The experimenters also varied the wording on some of the control questions, so that half the subjects received "nonexclusive" and half "exclusive" control questions.

In the pretest interview, the examiners followed the usual field procedure of reviewing the control questions with the subjects, and the questions were adjusted until they elicited a "no" response. The control question polygraph test then took place, with three or more charts obtained from each subject, although only the first three were used in the objective scoring. Immediately after testing, the first three charts obtained were scored blind on electrodermal response (EDR), respiration, and cardio measures. Later, an independent rater scored the tests, using the numerical scoring system devised by Barland and Raskin *(2 I).* The experimenters also used objective measurements of all physiological response measures with the aid of computers and persons who had no knowledge of the field evaluations or treatments administered. The experimenters used the deci-

sions made by the independent blind evaluator to assess the validity of the polygraph test. This was, however, equivalent to using the polygraph examiner's decision, because the independent rater and the examiner agreed on 100 percent of their decisions.

The results for both types of control questions combined (with an inconclusive zone used) were *80* percent correct, 10 percent incorrect, and 10 percent inconclusive. There were three false negatives (8 percent) and one false positive (2 percent). The accuracy of CQT differed depending on whether exclusive or inclusive control questions were used. When the exclusive control questions were used, 85 percent of the subjects were identified correctly, 5 percent incorrectly, and 10 percent inconclusively. Of the assessments of the 20 subjects in this condition, one (5 percent) was a false negative and there were no false positives. When nonexclusive control questions were used, 75 percent were correct, 15 percent incorrect, and 10 percent inconclusive. Of these 20, two (10 percent) were false negatives and one (5 percent) was a false positive. The evaluative scores for each physiological component were analyzed to determine if the scores differed between guilty and innocent subjects. Only the EDR and plethysmograph scores yielded significant differences.

Behavioral observations, by themselves, yielded a significant number of correct decisions, but this differed greatly between innocent and guilty subjects. Of the guilty subjects, 86 percent of the decisions made were correct (25 deceptive, 4 nondeceptive, and 1 inconclusive); however, only 48 percent of the innocent subjects were correctly identified (12 deceptive, 11 nondeceptive, 7 inconclusive). An objective quantitative analysis for each physiological measure was employed to determine if each was effective in discriminating between guilty and innocent. Most of the measures yielded significant discriminations, with the exception of a few of the cardiovascular measures.

Raskin and Hare

A special population, prisoners, especially relevant to the field use of the polygraph, was the focus of a study by Raskin and Hare (137). In their sample of 48 inmates of a Canadian prison, half were selected for high levels of psychopathy, and half for low levels. One purpose of their study was to investigate whether deceptive psychopaths could more easily escape detection than normal subjects (see ch. *6).* Overall, assessments of deception from the field evaluations from all charts were *88* percent correct, 4 percent wrong, and 8 percent inconclusive. There were only two errors, both false positives. No significant differences were found between psychopaths and nonpsychopaths, suggesting that a CQT polygraph examination is equally valid for both. Also, a quantitative analysis showed that all the physiological measures were significantly different between guilty and innocent subjects. Psychopathy did not obscure these differences and in some cases enhanced them.

Rovner, Raskin, and Kircher

Rovner, Raskin, and Kircher (143) studied the effect of information and practice on the accuracy of polygraph examinations. Seventy-two subjects recruited from the community took part in this mock crime experiment. One third of the subjects (12 innocent and 12 guilty) were given in-depth information about the polygraph and about countermeasures used to appear innocent (information condition). Another third received this information and underwent two practice polygraph examinations about which they received "feedback" (information and practice condition). The other third had no such intervention (standard). A blind field evaluation performed some time later produced the scores for decisions of guilt or innocence, and for analysis of the physiological responses. Accuracy for the standard group and the information group was identical: 88 percent correct, 4 percent incorrect, and 8 percent inconclusive. But accuracy for the information and practice condition *was* lower: *62.5* percent correct, 25 percent incorrect, and 12.5 percent inconclusive. There was one error in the standard group and one in the information group—both false positives. The six errors in the information and practice conditions were three false positives and three false negatives.

Kircher

Some of the latest work of the Utah laboratory explores the use of computers in the analysis of

polygraph recordings. Kircher (91a) compared the accuracy of a computer decisionmaking process to the accuracy of assessments of a field examiner. The computerized analysis cannot be included in the statistical analysis of this technical memorandum, because it is not presently a field scoring method, but the decisions of an independent evaluator who was used can be. This mock crime study followed the basic procedures of Podlesny and Raskin (127) with 100 subjects from the community. The accuracy of the original examiner was not reported though the results of an independent evaluator were. The independent evaluator, who numerically scored the charts blindly, correctly diagnosed 87 percent of the subjects; misdiagnosed 6 percent; and made a judgment of inconclusive on 7 percent. The six errors were evenly divided between three false negatives and three false positives. In comparison, different computer decision models, on the average, correctly identified 84.9 percent of subjects, misidentified 7.85 percent, and placed 7.2 percent in an inconclusive category.

Other Studies

A range of other studies has been conducted in recent years to evaluate aspects of polygraph test validity. Such studies usually manipulate one or two variables that are hypothesized to be important determinants of polygraph validity. For the most part, these experiments use procedures that are similar to Raskin's mock crime paradigm. Some of the discussion of the procedures in each study is omitted, because they closely follow this paradigm.

Dawson

Dawson (49), for example, focused on the effect of "cognitive countermeasures" on validity. His study was unique in that the subjects were actors trained in the Stanislavsky method of acting, which teaches actors to use their own experience to create emotional states appropriate for a role. Studying the attempts of "method" actors to foil the polygraph may help determine whether guilty subjects can be trained to use cognitive countermeasures to appear innocent (see ch. 6). Dawson was also interested in analyzing separately responses during two distinct phases of the questioning: while subjects listened to questions and while they responded.

Dawson's sample consisted of *24* student actors, half of whom were randomly assigned to the "guilty" group and half to the "innocent" group. They were instructed to use the Stanislavsky method to appear innocent on the polygraph examination. After the mock crime, four charts were obtained from ZOC control question test about the crime. On two of the charts, the subjects were instructed not to respond until they received a signal 8 seconds after a question. This served to separate responding associated with the questions from responding associated with answering. Numerical scoring based on Barland and Raskin's (21) system was done separately on three different types of physiological responses:

1. responses when the answers were immediate;
2. responses during the questions when the answers were delayed; and
3. responses during the answers when the answers were delayed.

Dawson found that the subjects' immediate physiological responses to the questions, whether they were answering immediately or not, led to decisions which were 88 percent correct, 8 percent incorrect, and 4 percent inconclusive (frequencies across two conditions were summed). The delayed answer response yielded a rate of 29 percent correct, 8 percent incorrect, and 62 percent inconclusive. The incorrect decisions made were entirely false positives. A quantitative analysis revealed that the EDR and cardiovascular measures differentiated significantly between innocent and guilty, but respiration did not. The major outcomes of this study suggested that the polygraph was not susceptible to cognitive countermeasures of the sort used by the actors and that scorable responses generally occur immediately after questions.

This experiment does not, however, test cognitive countermeasures in a situation in which the subjects know the essentials of CQT and apply cognitive countermeasures differentially to relevant and control questions. The average criminal subject is likely to attempt cognitive measures naively, but a sophisticated subject—perhaps the type more likely to appear in a national security investigation—may learn cognitive countermeas-

ures along with the knowledge of the control question or other technique.

Widacki and Horvath

Widacki and Horvath (207) designed an experiment to examine the polygraph's efficacy in comparison to other techniques in the mock investigation of a mock crime. They recruited *80* Polish student volunteers and had all of them provide writing specimens, photographs of themselves, and fingerprints. Subjects were then assigned to *20* groups of four subjects each. Within each group, one subject was randomly assigned to be the perpetrator, and the other three were innocent suspects. Each group was thus an "investigative case. " Because of this feature of the design, the decisions of guilty and innocent were not independent. Therefore, Widacki and Horvath's findings could not be included in the statistical analysis of the control question analogs and must be considered separately. A similar situation holds for Kubis' (93) mock crime experiment (see below).

The mock crime proceeded as follows: the guilty subject picked up a parcel from one of two persons acting as a "doorkeeper" of a building in the area. The perpetrator gave some experiment-related papers to the doorkeeper and then signed for the parcel. Thus, an eyewitness account (by the doorkeeper), fingerprints, and handwriting specimens were all available. Blind polygraph examinations then were conducted using the Reid control question method (including the examiners' behavioral observations of the subject). Analysis of the three other sources of evidence was carried out.

Widacki and Horwath found that the polygraph produced the most correct decisions (n =18), the fewest (along with handwriting) incorrect decisions (n= 1), and the fewest inconclusive decisions (n=l). Widacki and Horvath note, however, that a direct comparison of these four investigative methods may be invalid because the experimental procedures could not ensure a comparable level of quality of evidence for each method (e.g., fingerprints were not detectable in the majority of cases).

Because of its experimental design that had the examiner make decisions on four suspects as a group, the study produces data about the accuracy of the polygraph that is difficult to interpret. But it does shed light on the efficacy of the polygraph relative to other investigatory techniques that might be the alternative. Certainly, it is crucial in policymaking to judge the validity of the polygraph relative to other techniques that would be used in its stead. More research is needed in which the polygraph is compared to other investigatory techniques, and the quality of information across techniques is held constant. Such a comparative analog study would be especially valuable if it included different techniques used in investigations of Federal personnel, such as those reported in the Director of Central Intelligence (DCI) survey mentioned in chapter 4 ("administrative screening," "investigative interviews, " etc.).

Bradley and Janisse

Bradley and Janisse (35) studied the effects of two other variables hypothesized to influence the validity of the polygraph: the degree of threat involved in the punishment for being judged guilty, and successful demonstration to the subjects of the technique's accuracy. A mock crime was carried out using procedures similar to those used by Barland and Raskin (21). Subjects were also given a series of stimulation tests. Results of these tests were manipulated such that they made the polygraph test appear perfectly effective, partially effective, and ineffective. In addition, half the subjects were told they would receive a painful electric shock if found guilty, though no shock was ever given.

The degree of manipulated effectiveness had no direct effect on scores, but did tend to increase the accuracy of detection. Threat of punishment did not affect accuracy of detection, although it did have an overall effect on heart rate. EDR and heart rate change were significantly accurate in differentiating guilty and innocent, although another measure, pupil size change, was not.

Honts and Hodes

Two recent analog studies of the Backster ZOC method of testing (76,77) were conducted primari-

ly for the purpose of testing whether polygraph examiners could detect the use of physical countermeasures by subjects. In the first study, subjects were college students who received extra credit toward their final grades for their participation. "Guilty" suspects participated in a mock crime (theft of an examination); innocent suspects were only told of the theft. All subjects were motivated to produce truthful outcomes on the polygraph test by an offer of twice the number of credits if the examiner reported them as truthful.

In addition to participation in the mock crime, *24* of the guilty subjects participated in 15-minute training sessions in which they were told about the theory of CQT and shown how to use either tongue biting (12 subjects) or toe pressing (12 subjects) as countermeasures during presentation of the control questions. They were also instructed to try to relax as much as possible during presentation of the relevant questions.

The actual polygraph examinations took place a week after the theft and training sessions. All guilty subjects were instructed to have the "stolen" examinations with them, presumably to enhance subject involvement. Four charts were obtained from each subject using a standard Backster examination administered by an experienced polygraph examiner. The examiner was aware of the details of the experiment, including a knowledge of the base rates of guilt and the countermeasures that would be attempted, but was blind to the group assignment of individual subjects. At the end of each examination, the examiner made a yes/no decision regarding the subject's use of countermeasures. After all subjects had been tested, the original examiner made a decision as to deception by blindly evaluating the charts using the Backster numerical scoring technique, and made another decision about the use of countermeasures based on inspection of the charts. Charts were also examined and scored by a second examiner who was blind to all aspects of the experiment.

As shown in table *6,* while there was a low rate of false negatives *(5.5* percent), examiners were not able to make a decision on one-third of countermeasure and no countermeasure guilty subjects, and half of the innocent subjects. There was a 7 percent false positive rate. Examiners were not able to detect the use of countermeasures.

In their second experiment on countermeasures, Honts and Hodes used approximately the same procedures and subject pool, with the exception that subjects were asked to employ both countermeasures simultaneously, were given *30* minutes of training, including a practice session, and were asked to practice at home. A cardio cuff was added to the polygraph instrument, and a card test was conducted prior to the administration of the first test.

Overall, results of the second study replicated the first. The categorizations of the original examiner were 51 percent correct, 14 percent incorrect, and 35 percent inconclusive. Twenty-six percent of the countermeasure subjects compared to none of the guilty/no countermeasure subjects were incorrectly classified as truthful. Examiners were not able to detect successful countermeasure users.

Ginton, et al.

Only two CQT analog studies have used real, albeit small, crimes. Another study using a real crime tested the GKT technique and is discussed in the section on that technique. An experiment by Ginton, et al. *(62),* aimed to create even more verisimilitude than usual in the analog study. An auxiliary purpose was to test the relative effectiveness of behavioral observations, global evaluations, including behavioral observations, and numerical scoring based on the charts alone.

Subjects in Ginton, et al. 's, investigation were *21* Israeli policemen. They were given paper and pencil tests that were presented as required aptitude tests. Subjects were asked to score their own tests, which provided an opportunity to cheat, i.e., to revise their initial answers. The test answer sheets, however, were chemically treated so that cheating could be detected. Seven of the twenty-one subjects actually changed their initial answers. Later, subjects were told they were suspected of cheating, were offered an opportunity to take a polygraph examination, and were told their careers might depend on the outcome. Fifteen sub-

jects actually underwent the polygraph testing, only two of whom had actually cheated.

A CQT was administered, and each subject was evaluated by three polygraph experts who had conducted or witnessed the particular examination being evaluated. One examiner (an observer) relied on behavioral observation, another (a rater) used only the charts, and a third (the actual examiner) used both sources of information. The evaluations were made globally. Five other polygraph examiners evaluated the charts later using both the Utah group's scoring system (21) and global evaluations. The original three performed a second analysis in this way, too. Conclusions about this study are limited because of a large no-show rate among the guilty subjects. Both guilty subjects who took the test were correctly detected. However 15 percent of the noncheaters were incorrectly identified as deceptive.

Heckel, et al.

Another analog study (74) used a staged crime to investigate the differential accuracy of CQT with psychotic, neurotic, and normal subjects. Fifteen subjects (five from each of the above three groups) were given the opportunity to steal money from the wallet of an experimenter who was staging a session of psychological testing. The experimenter later alleged that *$20* had been stolen, and arranged for polygraph examinations of the 15 subjects by a field examiner. No money had actually been stolen, so the subjects were actually innocent. Four polygraph experts later rated the charts. Averaging the results for these independent evaluators, 11 of the subjects were correctly labeled innocent, 1 was called guilty, and 3 were placed in an inconclusive category. The one error and one inconclusive were with psychotic subjects, and the other two inconclusive were with neurotic subjects. Because only innocent subjects were included, a lambda was not calculated for this study.

Hammond

Hammond (64a) conducted a mock crime study to test the hypotheses that: 1) alcoholics would be less detectable than normal subjects, 2) psychopaths would be as detectable as normal subjects, and 3) student examiners would not be as accurate as an expert examiner. He was also interested in the overall value of polygraph examinations for forensic psychology. The subjects in Hammond's study were volunteers solicited through sign-up sheets in a college fraternity (normals), alcoholism treatment centers (alcoholics), and ex-offender programs (psychopaths) as well as through newspaper advertisements and other means. Psychological tests (e. g., subscales of the MMPI) as well as polygraph examinations were given to the subjects. The polygraph examinations were conducted by students near the end of their training at the Backster School of Lie Detection. Examiners used a version of Backster's control question technique, and Backster's numerical scoring system. Charts were scored using several levels of inconclusive zone by both the student examiners and an expert examiner who scored the charts blindly. Two polygraph charts, rather than the standared three, were conducted for each subject.

Table *6* shows the results of Hammond's study using the standard *8 inconclusive zone. As shown, approximately *72* percent of the guilty subjects and *40* percent of the innocent subjects were scored correctly. Neither alcoholics, normals, nor psychopaths showed differences in detectability. In addition, there were no differences between the numerical scores of the student examiners and the blind expert examiner. However, using the *8 cutoff, expert evaluators had more inconclusive (and fewer innocent "hits") than the student examiners. While Hammond concluded that his study supported the validity of polygraph testing, he believed that certain factors in his study could account for the failure to show differences by subject category. In particular, all subject groups actually turned out to be relatively heavy drinkers. Hammond also contended that overall accuracy rates would have been higher with more experienced polygraph examiners. He observed that the examiners in his study were unskilled at detecting countermeasures and at calibrating the polygraph instrument.

Szucko and Kleinmuntz

A somewhat different approach to assessing the validity of the polygraph was taken by Szucko

and Kleinmuntz (160). They directly compared the ability of polygraph examiners to assess deception against the ability of computers to do the same using a digitalized form of the same data. They had a sample of 30 psychology undergraduate volunteers and randomly assigned them to the guilty or innocent conditions. The mock crime involved the "theft" of a $5 bill. Polygraph tests were administered by four examiner-trainees from a polygraph firm near Szucko and Kleinmuntz's university. The recordings of the physiological measures were transformed into digital form for computer analysis.

Six experienced polygraph examiners independently evaluated the charts. No inconclusive category was allowed in the study. Digital polygraph data was evaluated by computer. A lens model equation drawn from studies of human judgment was used. The results of this analysis indicated that five of the six polygraph raters were able to detect deception significantly better than chance, but four of them also had fairly high rates of false positives. Szucko and Kleinmuntz estimate that the judges detected on the average 71 percent of guilty subjects, but also called half of the innocent subjects deceptive (false positive). Szucko and Kleinmuntz state that 80 percent of the protocols could be classified correctly using a purely statistical analysis, but they do not state the detection rate, false positive rate, and false negative rate of their statistical analysis.

Kircher and Raskin (91) contend on the other hand that evaluators using numerical evaluations can be "at least as accurate as those produced by any known statistical decision model and that the accuracies of both clinical and statistical methods exceed 90 percent ." Kircher and Raskin reanalyzed charts from the Rovner, et al. (143), study described above and used a lens model, similar to that employed by Szucko and Kleinmuntz. The studies, however, differed in a number of ways, which could account for the variation in their results. Probably the most important difference is that Kircher and Raskin's interpreters were trained in numerical scoring procedures (136), whereas interpreters in the Szucko and Kleinmuntz study used global evaluation procedures (139).

CONCEALED INFORMATION TESTS

Although the largest number of analog studies investigate CQT, several analog studies have examined the validity of the guilty knowledge test, one type of concealed information test. A search of the literature revealed no analog studies of the peak of tension test as a distinct technique.

Lykken

In one early investigation of GKT, Lykken (105) attempted to demonstrate that the detection of incriminating knowledge about a crime can be done more accurately than the detection of a lie about the crime. In Lykken's study, 49 male college students were randomly assigned to four categories of guilt in conducting two mock crimes. Subjects either committed a staged "theft," a staged "murder," both, or neither. An experimenter then conducted two GKT polygraph examinations with each subject, one for each crime.

Each test in Lykken's study *(105)* included six questions about details related to the "murder" situation and "theft" situation (e.g., asking the subject to identify an object present in the "murder" room), To make subjects anxious about the accuracy of their responses during the examination, they were told they would be given an electric shock if the examiner felt their responses indicated guilt; in fact, subjects received an electric shock after every question. The relevant alternative in each question was randomly varied among an average of five possibilities. If the question about the relevant detail produced the EDR with the greatest amplitude, it received a score of '2. " If it was the second largest in amplitude, it received a "1. " A perfect guilty score on each test was *"12,"* and a perfect innocent score was "O. " A score of seven or greater was categorized as guilty for the purpose of analysis, and a score of six or less was categorized as innocent. The guil-

ty knowledge test was accurate to a significant degree in identifying subjects who committed both, either, or neither of the crimes. On the basis of this experiment, Lykken argued that GKT, with some refinements, could be applicable in criminal investigations.

Davidson

Other researchers have used Lykken's GKT paradigm to explore further its validity as a polygraph examination technique. Davidson (46) examined the GKT's validity under conditions that varied motivation level and that he claimed were, in general, more "ego-involving" for subjects. In Dawson's study, 48 college students were recruited and assigned randomly to 12 groups of 4. Three of the four were instructed to attempt to commit specific mock murders, and the fourth served as a control. The mock crimes were arranged such that one subject would "commit" the crime, one would try to fail, one was motivated but never had the opportunity, and one (the control) had no knowledge of the crime. Half of the subjects who "committed" the murders received a large amount of money *($25* to $50) and half received a small sum ($10 to $1). The different amounts were presumed to create a different level of motivation in the subjects. The subjects then were examined with the use of GKT. Six multiple-choice questions with five alternatives were presented to the subjects, and the EDR was recorded. The scoring method followed Lykken's (105) exactly (see above). Using a weighted average, 98 percent of the classifications were correct against a chance level of 25 percent. The only error was one false negative.

Podlesny and Raskin

Podlesny and Raskin (127) included GKT in their study of a variety of polygraph techniques and physiological measures. Their experiment was unique in that it employed GKT in the same context as CQT (see above). Thus, they were able to compare the accuracy rates of the two techniques, although they claimed that a different statistical comparison was impossible because the two techniques use very different methods to assess guilt. Podlesny and Raskin also were the first to test GKT with physiological measures other than EDR. To make assessments of guilt, they used the traditional polygraph respiration and cardio measures, and another vascular measure that was a composite of finger blood volume and finger blood amplitude. This latter measure was recorded by the photoplethysmograph mentioned above. In addition, Podlesny and Raskin performed a quantitative analysis of differences between guilty and innocent subjects on several other physiological measures.

GKT was conducted after the same mock theft Podlesny and Raskin (127) used to study CQT. Twenty subjects (10 guilty and 10 innocent) were examined with GKT, which included five questions with six alternatives each. The relevant alternatives were placed among the other alternatives in a "pseudo-random" order (127). The GKT charts were scored by the same method used by Lykken (105) and Davidson (46). Podlesny and Raskin also scored the charts in another way, with the addition of an inconclusive zone of scores five or six. This scoring system for assessing guilt was used with the photoplethysmograph, respiration, and cardio measure as well as EDR. Their findings were that GKT with EDR was correct for 90 percent of the subjects and incorrect for 10 percent, all false negatives. Using an inconclusive zone did not add significantly to the accuracy of the technique, however: 80 percent of assessments were correct, 10 percent incorrect (all false negatives), and 10 percent inconclusive.

Giesen and Rollison

Giesen and Rollison (61) studied the effects on GKT of the subjects' trait anxiety levels and of the possibility that crime-related details could be relevant to innocent subjects because of associations unrelated to the crime. Trait anxiety is anxiety that is characteristic of one's personality and would be relatively stable over time. Both trait anxiety and "innocent associations" could conceivably confound the detection of guilt with GKT.

Giesen and Rollison selected 40 female undergraduates who responded positively to a questionnaire item on "palmar sweating. " EDR is related

25-290 0 - 83 - 6 : QL 3

to sweating. Thus, this sample may have tended to produce higher EDRs than the norm. This group was divided into two groups of 20: those who scored high on a questionnaire measure of anxiety (Lykken's activity preference questionnaire) and those who scored low. Ten subjects in each group were then assigned to the guilty knowledge condition, and to the "innocent associations" condition. The guilty subjects were told to pretend to be secret agents who had committed a murder. They read a narrative about the crime, and role-played the act of burning an incriminating picture. Innocent subjects also played secret agents, but read a narrative containing several details (e.g., how much money was involved), which in the guilty condition were related to the crime. They had, therefore, as much exposure to this information as the guilty subjects, but in an innocent context. Using GKT with EDR, experimenters asked subjects eight crime-related questions, each with five alternatives. Those details common to both conditions were used as the crime-relevant items in GKT questions. Scoring followed Lykken's *(105)* method.

Giesen and Rollison found that GKT was highly accurate, correctly classifying all of the innocent subjects and detecting all but one of the guilty subjects (an average of *97.5* percent correct). In addition, they found that the EDR measure was significantly different between guilty and innocent subjects. Trait anxiety level had no effect on EDR by itself, but the more anxious subjects in the guilty condition had significantly greater EDR than the less anxious, especially in response to the relevant items. These findings would suggest that anxiety alone does not confound GKT results, but anxiety in guilty subjects might indeed augment the accuracy of the technique. The study also suggests that GKT may be accurate even when innocent subjects have greater associations with crime-relevant items than with neutral items. This finding, however, must be tempered by the fact that the entire sample was selected for their tendency for palmar sweating under stress and, thus, may be unrepresentative of polygraph subjects in general.

Balloun and Holmes

Balloun and Holmes (12) used GKT to detect guilt in a "real" crime arranged by the experimenters. They were also interested in the effect of psychopathy and of repeated examinations on the accuracy of GKT. They selected 18 male college students with high scores on the psychopathic deviate (Pal) scale of the Minnesota Multiphasic Personality Inventory (MMPI) and 16 with low scores. The Pd scale was originally designed to make the diagnosis of psychopathic personality and was used as a scale to measure relative "amounts" of psychopathy. The experimenters acknowledge, however, that the Pd scale may be an inadequate measure of this diagnosis. These subjects took a fake intelligence test with two other students (actually confederates of the examiner). The confederates urged subjects to cheat and supplied test answers to those who were willing. Eighteen of the thirty four students cheated. Later, the subjects underwent a polygraph examination using GKT. They were reminded that cheating on exams could lead to academic dismissal, and that the experimenters knew that some had cheated on the "intelligence test. " Information from the intelligence tests that only the cheaters would know served as the incriminating details on GKT. Another GKT with the same content, but a different order of questions was then administered to see if the subjects would adapt to GKT and, thus, reduce its accuracy.

Balloun and Holmes scored GKT using Lykken's *(105)* method with three physiological measures (EDR, heart rate, and finger pulse volume), but only EDR produced significant results. On the first test, guilty subjects scored significantly higher and were detected with significant accuracy. However, on the second test, though the guilty subjects had significantly greater scores, they were not great enough for significantly accurate detection of guilt at the criterion level *(5.5* out of 10) used. There was no difference between the high and low Pd subjects on either administration of GKT.

Bradley and Janisse

In their study of the influence of threat and demonstrations of accuracy on the polygraph examination (see above), Bradley and Janisse (35) also tested the 192 subjects with the GKT after the CQT had been conducted. The questions concerned four relevant details. They were scored using the Lykken *(105)* method. With EDR data, the GKT classified an average of 74 percent of subjects correctly, and *26* percent incorrectly with 11 false positives and 39 false negatives. With the measure of heart rate change, the GKT categorized *63.5* percent of subjects correctly and *36.5* incorrectly, with 17 false positives and 53 false negatives, Neither the degree of threat nor the demonstrated effectiveness of the polygraph test had a significant effect on the discrimination between deceptive and truthful subjects.

Timm

Timm (163) examined the effect of the administration of a placebo on the validity of GKT. Also included in the experiment was an investigation of the effect on GKT accuracy of differential feedback from the stimulation test. In the experiment all *270* college student subjects committed a mock crime. There were no "innocent" subjects. Before the mock crime, subjects were either: 1) given a placebo and told it would help them "beat" the test; 2) given a placebo and told it would make it more difficult to deceive the examiner; or *3)* not given a placebo. The stimulation or number test was arranged to produce three different feedback conditions. One-third of the subjects' numbers were detected, one-third were not, and one-third did not receive the results of the stimulation test. After the GKT was conducted on each subject, charts were scored according to the Lykken *(105)* method. Adequate charts were obtained for *237* subjects. Of these subjects, *70.4* to *80.8* percent of them produced scores indicative of guilt, depending on how conservative a cutoff point for the score was used. Neither the placebo condition nor the feedback condition produced a significant effect on detection ability. Because of the absence of "innocent" subjects in this study (i. e., a base rate of guilty of 100 percent), the study tells us nothing about the accuracy of GKT with the innocent subjects. And even the results with guilty subjects are difficult to interpret when there is no comparison to results with innocent subjects. Also, without innocent subjects, a lambda is impossible to calculate.

PREEMPLOYMENT SCREENING

Despite its widespread use in the field, there are few analog studies of the preemployment screening polygraph examination. The two that are known to employ post-1960 polygraph screening techniques are reviewed. Correa and Adams (43) conducted an analog investigation of this type of examination with *40* undergraduate subjects. Barland *(16)* conducted an analog study with Federal Government personnel.

Correa and Adams

Like the usual preemployment screening test, the examination in Correa and Adams' study included a number of relevant questions. Subjects were interviewed prior to the polygraph examination and completed a questionnaire about their background. Half the group was instructed to lie to nine relevant questions and half to tell the truth. The polygraph test was conducted, and three charts of 32 questions each were recorded. Most of the relevant questions concerned information from the questionnaire, but also included were three questions about events staged by the researcher in the initial interview (e.g., giving the subject a glass of water). These latter questions served as a check on the honesty of subjects in completing the questionnaire, and were considered relevant questions in the evaluation of deception or nondeception. The examiner subjectively made assessments of veracity based on the polygraph recordings. When questions about the staged events and the application were diagnosed by the examiner, all *40* of the subjects were correctly identified as being deceptive *or* truthful.

Correa and Adams conducted a question-by-question analysis of the charts of deceptive subjects. A mean of 75 percent of the relevant items from the screening application were correctly classified, and a mean of 25 percent were incorrectly classified. When change scores were calculated for each physiological response, all physiological measures (EDR, respiration, cardiovascular) significantly discriminated truthful from deceptive subjects. Correa and Adams suggest that these findings provide evidence for the validity of prescreening polygraph examinations. There are, however, a number of problems with the Correa and Adams' study that may compromise its validity. Several features of the experiment are probably highly unrepresentative of or unrelated to field preemployment polygraph examinations: the length of the interview *(96* questions); the number of deceptive responses subjects made (9); and the inclusion of questions about the experiment itself. Furthermore, the experimenters fail to discuss the criteria by which the assessments of veracity were made, so it is difficult to ascertain whether these assessments correspond to field assessments.

Barland

The Barland (16) study is important for several reasons. One, subjects were actual military personnel who in Barland's opinion might be the types screened for counterintelligence purposes. All subjects were assigned to intelligence duties. It is, thus, unique in being the only validity study of preemployment screening in an intelligence context. However, because it did not ask any questions related to security interests, it cannot be considered a full analog to field situations.

Second, it tested the validity of a type of CQT, the directed lie control question (DLCQ) technique, in a screening situation. DLCQ is part of a counterintelligence screening test developed by Army Intelligence examiners in 1971. During the pretest phase of this technique, subjects typically answer "yes" to certain questions. When they answer yes, the examiner instructs them that when they are asked such questions during the actual polygraph examination, they should respond with a "no" rather than a "yes. " Thus, they are directed to lie, and their lies to these questions constitute the control questions against which responses to relevant questions are compared. DLCQ differs from the control question discussed previously (see ch. 2). With the DLCQ technique, the control questions are not designed to provoke the subject to lie or be concerned about the telling the truth. The "lies" do not constitute deception since the examiner instructs the subject to tell lies that they both know are false. However, the directed lies are believed to generate concern in innocent subjects because the subjects are told that to appear nondeceptive on the rest of the examination, they must appear deceptive on the directed questions.

The question of whether CQT can be used outside of specific issue investigations (e. g., in preemployment or periodic screening) is controversial. It is difficult to construct standard control questions when much of a person's past is irrelevant to the purpose of the examination, since past misdeeds (i. e., other than the specific issue being investigated) typically comprise the subject area of control questions.

In this 1981 study, Barland solicited volunteers from the military intelligence community. Subjects were told the purpose of the study and that testing would be limited to the subject's date of birth, place of birth, education, employment, and residences (these were the relevant items), and that some subjects would be instructed to furnish the examiner with false information. Approximately half the subjects were instructed to lie to one of the relevant items; these subjects were offered a $20 reward if they could appear truthful on the polygraph examination. Unlike the data in the Correa and Adams' study, the experimenter was able to check the information given by the subjects against data obtained from background investigations. The three polygraph examiners in the study had 3, *6,* and 9 years of polygraph experience and had been trained at the U.S. Army Military Police School (USAMPS) polygraph course,

Examiners used three methods of chart interpretation: zone of comparison, greatest control method, and relevant-irrelevant method. As explained in chapter 2, in the zone method, relevant questions are evaluated against the larger of either

control question response in a zone. In Barland's (16) zone method, each physiological measure for each relevant /irrelevant control question pair was rated on a point scale using interpretive criteria taught at USAMPS. In the relevant-irrelevant method of interpretation, each relevant question was evaluated without making specific reference to the control question nearest it; emphasis "was placed on the size and consistency of reactions at the relevant questions" and scored globally rather than numerically. The "greatest control" method consisted of evaluating all five relevant questions against the single control question on that chart which had the largest overall reaction. In addition to the comparisons of the three chart interpretation methods, charts *were* analyzed globally and on a question-by-question basis.

In the global method of analysis, subjects were categorized as either deception indicated, no deception indicated, or inconclusive on the basis of appearing deceptive to any of the relevant questions, That is, if a subject was in fact deceptive to any relevant question, and he reacted deceptively to any of the questions, it was considered a hit even though the examiner may have misidentified which relevant question the subject was deceptive to. Using this method of assessing deceptiveness, the three methods of chart interpretation achieved the following results:

Zone:
- 62 percent correct identification of truthful subjects;
- 19 percent incorrect;
- 19 percent inconclusive;
- 70 percent correct identification of deceptive subjects;
- 17 percent incorrect;
- 13 percent inconclusive.

Greatest control:
- 77 percent correct identification of truthful subjects;
- 15 percent incorrect;
- 8 percent inconclusive.
- 50 percent correct identification of deceptive subjects;
- 23 percent incorrect;
- 27 percent inconclusive.

Relevant-irrelevant:
- 73 percent correct identification of truthful subjects;
- 23 percent incorrect;
- 4 percent inconclusive.
- 80 percent correct identification of deceptive subjects;
- 13 percent incorrect;
- 7 percent inconclusive.

Presumably, the correct identification rates would be lower if only those cases in which the truly deceptive relevant response was counted as a "hit. " To test this hypothesis, the authors conducted a question-by-question analysis. In this method, identification of truthful responses increased but identification of deceptive responses declined quite a bit, Using the zone technique, 77 percent of the truthful questions and only 57 percent of the deceptive questions were correctly identified. With the greatest control scoring method, *85* percent of truthful responses and less than half (43 percent) of deceptive questions were correctly identified, The R/I scoring technique showed the best results. With this method, *88* percent of the truthful subjects and *67* percent of deceptive questions were correctly identified (although global results were better with the R/I technique). This interpretation should be modified by the fact that each examiner used all three scoring techniques and the R/I technique was the last one used. Thus, the interpreter had the benefit of his previous judgments. The results of a blind analysis using other interpreters were not ready to be reported by Barland at the time his 1981 report was submitted.

The results of the Barland study raise serious questions about the usefulness of directed lie control questions in screening procedures as well as, in general, the validity of polygraph testing for preemployment and counterintelligence purposes, especially if used alone. Of course, the limitations of analog studies should be taken into consideration. Because of these limitations, Barland considers his results a "worst case" scenario. Finally, interpretations must depend on the false positive and false negative rates which are deemed acceptable for particular purposes.

FINDINGS

Separate statistical analyses were performed for the guilty knowledge and control question analog studies. The following data for the analog studies discussed above were reviewed:

- percentage of guilty subjects judged deceptive;
- percentage of guilty subjects judged nondeceptive (false negatives);
- percentage of guilty subjects judged inconclusive;
- percentage of innocent subjects judged deceptive (false positives);
- percentage of innocent subjects judged truthful; and
- percentage of innocent subjects judged inconclusive.

Also, as with the field studies, an index of predictive association (lambda) was calculated. The results (see tables 8 and 9) indicate that the control question test provides a 43-percent improvement in prediction over the base rate for these analog studies, and the guilty knowledge test a 70-percent improvement in prediction over the base rates. Because the studies differed so much, lambdas were calculated separately for each study. As shown in tables 6 and 7, individual lambdas ranged from zero to 83 percent for the CQT studies and 38 to 95 percent for the GKT studies (see ch. 4). These figures should be interpreted with caution as in real life the base rate of guilt will vary considerably from approximately 50/50 distributions in laboratory experiments. Thus, it is difficult to draw unqualified conclusions from the analog studies given the wide variety of designs used.

The false negative rate for the analog studies of CQT technique ranged from O to *29* percent. Inconclusive ranged from O to *44* percent for guilty subjects and from O to 53 percent for innocent subjects. There is a wide range of false positives (4 to 51 percent). Global evaluations by the examiners, field scoring techniques, and purely statistical analyses of the data all seem to produce high detection rates in most studies. One exception is Kleinmuntz and Szucko's *(92)* study, which found the validity coefficients of polygraph examiners' judgments markedly inferior to a purely statistical analysis of the charts. However, it is unclear how comparable their method of measuring validity is to the usual method of using an accuracy rate, and it is also not clear how applicable the lens model they use is to the question of the validity of the polygraph.

Another exception is Ginton, et al.'s, study *(62),* in which field numerical scoring was found to be inferior to the global evaluation method in detecting deception. However, the examiners in that study were Israeli polygraph professionals who may characteristically use a global method of assessment, and who may have been unfamiliar with the Utah numerical scoring system.

Accuracy of detection differed sizably between control question analog studies using students as subjects (Barland and Raskin, Bradley and Janisse, Szucko and Kleinmuntz; Widacki and Horvath is excluded as discussed above) and other control question analog studies (Podlesny and Raskin, Raskin and Hare, Rovner, et al., Dawson, Ginton, et al.). Experiments using students had lower percentages of correct decisions for both guilty and innocent, and more false negatives and false positives. Given the small number of studies in each category when the studies are divided in this way, it is unclear whether this difference is attributable to the nature of the subjects (student v. nonstudent) or other characteristics of these experiments.

As shown in tables 8 and 9, GKT analog studies detected a slightly lower average percentage of the guilty subjects than the CQT analog studies. They also had a relatively higher proportion of false negatives but a lower rate of false positives. It should be noted, however, that GKT was not assessed under conditions that deviated as much from the ideal as the control question test deviated. Nor were there as many studies testing GKT as CQT. This suggests that the confidence one can have in the GKT findings is, in general, less than the confidence one can have in the CQT findings.

In summary, there exists a number of studies of CQT; a smaller number of the concealed information test, all using GKT; and only two studies

of the preemployment screening interview, one of them with Government personnel. The analog studies systematically explored many of the technical variables associated with the polygraph (cf. the Utah group's studies of CQT), and also studied the effect of several situational variables on the validity of the polygraph. The control question test was found to detect guilty subjects with a relatively high degree of accuracy, but also to be subject to false positive errors. There was a large amount of variability among the control question analogs, especially the more they diverged in technique-from the field method. The guilty knowledge test had a slightly lower rate of detection of guilt, more false negatives, but fewer false positives,