Chapter 4

Review and Analysis of Polygraph Field Studies

Chapter 4

Review and Analysis of Polygraph Field Studies

INTRODUCTION

As noted in the discussion of previous scientific reviews of polygraph validity, considerable disagreement exists among reviewers as to which field studies and what kinds of evidence constitute acceptable tests of validity. This chapter presents the results of a systematic analysis of existing field studies of polygraph testing in order to make an independent assessment of validity. Field studies investigate actual polygraph examinations and constitute the most direct evidence for polygraph test validity *(27).* Both quantitative and qualitative techniques are utilized in order to make an overall assessment of existing evidence (63,125,142).

The goal of this analysis is to synthesize available research. Almost all of the available field evidence comes from cases involving specific-incident criminal investigations using the control question technique (CQT). This is an important limitation. Because a systematic review helps to identify this kind of problem, researchers and policy makers have a better basis on which to determine what, if any, additional studies are necessary. Also, the analysis aids understanding of which question techniques, test purposes, question designs, and scoring techniques have been studied and which may require further research. The analysis is designed to address many of the problems associated with qualitative or "literary" reviews of the research literature previously discussed. In particular, the analysis makes explicit the criteria used for both study selection and data analysis (63,125,142).

STUDY SELECTION

Studies were considered field studies of validity if their sample consisted of actual instances of polygraph examinations conducted by professional polygraph examiners, used field-tested polygraph techniques, and used some independent criterion to assess actual guilt or innocence. Although ground truth can probably never be known in an absolute sense, studies can be considered studies of validity only if they included some adequately described and systematically determined criterion of "truth" (e. g., panel decision, judicial outcome, confession). Studies in which judgments of one set of polygraphers are correlated with anothers' with no independent criterion of guilt or innocence are, in effect, reliability studies. Such studies have been excluded from the primary analysis reported here. Reports of unsystematically collected cases from police agencies and other organizations, in which the criteria for verification are unclear or unsystematic, have also been excluded.

The population of field studies considered for the present analysis was, in general, taken from those studies referred to in existing reviews of the scientific literature (see ch. 3). In addition, researchers active in the field of polygraph research were contacted and asked to supply the names and publication information of any additional recent studies. A bibliography provided by the American Polygraph Association *(9)* was also searched for references to field studies of validity. The 10 studies finally included (and listed in table 3) in the analysis are: Barland and Raskin (22), Bersh *(29),* Davidson (47), Horvath (82), Horvath and Reid (84), Hunter and Ash **(85),** Kleinmuntz and Szucko (92), Raskin (133), Slowick and Buckley (155), and Wicklander and Hunter (205). The fol-

Table 3.—Characteristics of Field Studies

Study	Type of validity affected: Criterion	Basis of Examiner Decision	Types of cases[a]
Bersh	Panel of legal professionals' assessment of investigative files	Original examiners' decisions	Criminal investigations/military personnel
Barland and Raskin	Panel of legal professionals' assessment of investigative files	Original examiners' decisions	Sex crimes, drug crimes, crimes of violence, crimes of financial gain, other crimes [b]
Barland and Raskin[c]	Panel	Blind evaluation	Sex crimes, drug crimes, crimes of violence, crimes of financial gain, other crimes
Barland and Raskin	Judicial outcome	Original decision	Sex crimes, drug crimes, crimes of violence, crimes of financial gain, other crimes
Barland and Raskin[c]	Judicial outcome	Blind evaluation	Sex crimes, drug crimes, crimes of violence, crimes of financial gain, other crimes
Raskin	Confession	Blind evaluation	Sex crimes, drug crimes, crimes of violence, crimes of financial gain, other crimes
Horvath and Reid	Confession	Blind evaluation	Theft, sexual misconduct, sabotage, bribery, criminal damage to property
Hunter and Ash	Confession	Blind evaluation	Theft, official misconduct, brutality, sexual assaults, homicide
Slowick and Buckley	Confession	Blind evaluation	Theft, industrial sabotage, drug abuse, rape
Wicklander and Hunter	Confession	Blind evaluation	Homicide, sexual assault, theft, official misconduct
Horvath	Confession	Blind evaluation	Crimes against persons, crimes against property
Davidson	Confession	Blind evaluation	Crimes against property/military personnel
Kleinmuntz and Szucko	Confession	Blind evaluation	Theft

~All studies use some version of control question technique
Only 77 of 92 cases were analyzed as to type of crime
~Not included in the analysis for reasons discussed in the text
Wicklander and Hunter also included an evaluation in which evaluators were given additional case material

lowing sections briefly describe the studies excluded from the analysis and the kinds of studies included in the analysis.

Studies Excluded

Not all studies referred to as field studies or actual criminal investigations by other reviewers are included in the present analysis. A comparison of studies shown in table 2 and the 10 studies included in the present analysis indicates that eight studies included by one or another of the reviewers are not included. The excluded studies are Bitterman and Marcuse *(30),* Ben-Ishai *(26),* two analyses reported in Raskin (133), Edwards **(52),** Elaad and Schahar (54), Peters (124), and Widacki *(206).* One study, Kleinmutz and Szucko *(92),* not included by various reviewers (because of its recent publication) has been included here. In addition, a number of studies included by Abrams (l), not shown in table 2, are also excluded from the present analysis. Many of the studies Abrams cited are excluded by later reviewers (e. g., Horvath (81)) because they are not actual validity studies (and did not use external criteria of "guilt/ innocence," e.g., MacNitt (113)), they did not use appropriate polygraphic instrumentation (e.g., Summers; see Abrams (1)), or did not use testing procedures common today (e.g., Lyon (111)). Other studies used by Abrams, but excluded from the present analysis, were unverified self-reports published in popular magazines (e.g.,

McEvoy (116)), or surveys of attitudes towards validity of the polygraph (e.g., Cureton (44)).

The Bitterman and Marcuse **(30)** study was excluded because, as pointed out by Lykken (108) and Horvath (81), among others, studies of single crimes for which there is only one possible guilty person raises the probability of accurate deception, regardless of method used, to a Ievel too high for the study to provide valid information, To give an extreme example, if there is one guilty suspect among 100 examined, making an a priori decision to calI them all innocent yields a 99-percent accuracy rate. In addition, Bitterman and Marcuse did not meet present criteria for field studies because the polygraphers were not professional examiners (they were psychology professors who had read books and articles about the polygraph technique), and they did not use field-tested measures of physiological response.

Ben-Ishai's *(26)* paper reports on two studies, both of which were excluded. One consisted of blind evaluations by Ben-Ishai of 10 polygraph charts. It is more accurately described as a study of reliability. The other used a single psychologist's (Ben-Ishai's) judgments of guilt or innocence based on investigative files as the criterion by which to judge polygraph accuracy. It is difficult to justify use of the judgment of a single psychologist as an adequate criterion of ground truth. Likewise, the information used to establish ground truth for the Elaad, Peters, and Widacki reports is not systematically collected and is inadequately described. These studies are more accurately described as a set of anecdotal reports. They use sampIes of cases collected from police files which are described as having been verified, sometimes by judicial outcome (Widacki), in others by confession (Elaad), and in the Edwards study, by "independent means. "

A final set of studies excluded are two of the three studies by Raskin (133). One analysis was directed primarily at an assessment of whether polygraph examinations are more favorable to defendants when conducted by polygraph examiners chosen by defense attorneys than when they are conducted by examiners chosen by prosecutors (the so-called "friendly polygrapher" hypothesis). The purpose of the second analysis was to discover the source of decision errors; these findings are discussed in chapter 6. The Raskin study included in the present analysis (133) was conducted with only the 16 cases from Barland and Raskin's (22) sample able to be verified by confession.

Studies Included

The field studies included are listed in table 3 in terms of the criterion used, the type of initial examiner decision, and the types of cases selected. These characteristics of studies relate to criterion, construct, and external validity, respectively.

The criterion dimension refers to the operationalization of ground truth used in a study. In one type of validity study, polygraphers' original decisions are compared against a criterion of ground truth established by a panel of experts (e.g., lawyers and judges). The panel makes their judgment on the basis of information in an investigative file, from which polygraph results are excluded. In another type of field study, a second set of examiners evaluates charts taken from a file. In most cases, the evaluation is "blind;" i.e., the examiner/evaluator does not know the original examiner's decision, the disposition of the case, nor any other information about the subject. In this situation, the original decisions have been verified by confession of the guilty party. Verification by confession is used as the ground truth criterion. In the third, and the least common type of field study, original examiners' decisions (the construct validity component) are judged against guilt or innocence established by judicial outcome, which is the ground truth criterion.

Researchers disagree about whether blind evaluations of polygraph charts or the decisions of the original examiners constitute true tests of polygraph validity. Whether one uses examiner decisions or physiological recordings depends on whether one is testing examiner decisionmaking or physiological arousal in response to certain questions. Blind evaluations of charts are probably less useful as research evidence because, in the typical examination situation, the decision as to suspects' deception is made by the original examiner and not by a blind evaluator. Even when examinations are subject to review (e. g., quality

control procedures used by the Department of Defense (DOD)), final decisions are still based on review of all information. Although a blind analysis is the first task of the quality control office, such quality control reviews do not fully control for the impact of a variety of factors, such as interpersonal expectancy effects which would still be reflected in the original polygraph charts. Interpersonal expectancy effects (141) refer to the possibility that an examiner's preexamination decision concerning guilt or innocence affects construction of examination questions or the psychological state of the suspect. Either of these could affect a suspect's physiological responses. Therefore, in studies for which results of both original examinations and blind evaluations were included, as in Barland and Raskin (22), the present analysis uses results of the original examinations instead of those for blind evaluations. It should be noted, however, that in these cases it is difficult to determine to what extent the decisions are based on the charts and to what extent they are based on interaction with the suspect (see 27,92).

Operationalizations of ground truth (the criterion component of validity) are also problematic. Studies using panel decisions have been referred to as the only valid field research on the validity of examiners' decisions (81), yet there is no way to know whether panel decisions based on investigative files are, in fact, correct. Raskin (136) notes some of the problems with using judicial outcomes and other criminal justice system resolutions (dismissals, guilty pleas) as criteria for validity. Cases may be dismissed for lack of sufficient evidence rather than actual innocence. If a jury acquits a defendant, it is not possible to determine the extent to which the jury felt that the defendant was actually innocent or whether they felt that there **was** not enough evidence to meet the standard of "guilty beyond a reasonable doubt ." Many guilty pleas are actually confessions of guilty to (lesser) crimes; as Raskin notes, it is difficult to interpret the meaning of such pleadings in regard to guilt on the original charge. The result is that, using criminal justice system outcomes, polygraph examinations may appear to have a high number of false positives (in the case of acquittals), or false negatives (in the case of dismissals).

The use of confessions, the most frequently used criterion of ground truth, is problematic in three ways:

1. confessions, themselves, are not always valid;
2. if the confession occurs prior to or during a polygraph examination, it cannot be considered an independent measure of guilt; and
3. those who confess may be a select sample of subjects, as discussed further below.

In addition to the above problems, studies differ in the adequacy of their research design. The most serious problems concern sampling. In most reported studies, neither cases, examiners nor evaluators were selected randomly. In some studies (e. g., 22,84), the cases of only one examiner are sampled. Nonrandom selection leaves open the possibility that the studies are not investigating "polygraph testing" in general, but instead only a subgroup of practitioners or testing techniques. When random sampling is used (as in Bersh (29)), high rejection rates of cases selected for analysis create other sample bias problems.

Some sample selectivity of unknown magnitude and importance occurs when confessions are used as a criterion. Studies using confessions may be using only a select sample of examinations. The magnitude of this problem is illustrated by the fact that in the sample of 92 cases obtained by Barland (22,133) only 16 were able to be verified by confession (132).

To summarize, because of problems in operationalizing important components of validity, none of the field studies of validity can be taken by itself as an indication of polygraph testing validity, In addition, because of the different operationalizations of construct and criterion validity and variations in research design, the studies are not strictly comparable with each other, These studies, however, constitute the most direct evidence for validity currently available and are analyzed as a group in order to assess the current state of knowledge about polygraph testing,

CODING

In order to conduct the present analysis, each field study was coded for a number of variables which had either been referred to as important factors in previous reviews of the literature, or which were deemed relevant to the various components of validity described in chapter 3. If the needed information was not available from the studies as published, the study author(s) were contacted and asked to supply the information. Appendix C lists the coding categories including relevant validity components (panel decision or judicial outcomes; confession), as well as design information (sample selection, attrition rate, examiner/evaluators' knowledge of base rate of guilt). All codings were made by two reviewers and each instance of disagreement over coding was resolved before analysis.

Data were coded directly from information provided within the study report or from information directly provided by the authors, with the exception of one variable. The exception was the coding category "objectivity of ratings, " which required that the coder make a judgment from high objectivity to low objectivity. Scoring was judged high if some actual standardized measurement (e. g., using a ruler) was taken of the physiological recordings on the polygraph charts. A rating of medium was given if numerical scores were assigned to subjective assessments of suspects' guilt or innocence (see, e.g., *22,92),* low if ratings of deceptive or nondeceptive were based on global assessments of charts only, and very low if decisions were based on charts plus other available information (in particular, observation and interaction with the subject). Objectivity ratings were made both for the original examiners' judgments and the blind evaluators or judges.

Finally, six categories of outcome data from each study were recorded:

1. guilty/deceptive subjects judged correctly;
2. guilty/deceptive subjects judged incorrectly (i.e., judged nondeceptive);
3. guilty/deceptive suspects judged inconclusive;
4. innocent/nondeceptive subjects judged correctly;
5. innocent/nondeceptive subjects judged incorrectly (i. e., deceptive); and
6. innocent/nondeceptive subjects judged inconclusive.

Categories 2 and 5 are the false negative and false positive rates, respectively.

FINDINGS AND DISCUSSION

Three questions are of particular importance to an assessment of polygraph validity useful to policymakers:

1. Are polygraph examinations valid?
2. Given the wide range of outcomes reported across studies, what accounts for their variability?
3. How generalizable are the results of studies to the current and proposed uses for national security purposes?

i o answer the first question, data from the available field studies were analyzed to ascertain whether polygraph examination accurately differentiate deceptive suspects from nondeceptive subjects. For this analysis, the outcome frequencies for each category were converted to percentages, and average percentages within each category were calculated. A measure of predictive association (lambda, see *64,73)* was also calculated, although the use of a single measure is very limited due to the wide variability in study design.

The $lambda_b$ index shows the proportional reduction in the probability of error in predicting one category (in this case, deception) when a second category (in this case, polygraph examination results) is known. If the information about the second category does not reduce the probability of error in predicting the first category at all, the index is zero, and one can say that there is no predictive association. On the other hand,

if the index is 1.00, no error is made in predicting one category from another, and there is complete predictive association. Essentially, lambda provides an index that translates to the percent improvement over the base rate and indicates the percent improvement in prediction when the polygraph examinations are considered versus no further information. There is almost no direct research on the percent improvement of the polygraph over other forms of investigation (cf. 207). The results of this analysis of predictive association are shown in tables 4 and 5. The average lambda~ across studies is *0.65,* which means that, on the average in these field studies, the polygraph diagnosis reduced *65* percent of the error of chance prediction. The lambdab for individual studies ranged from *0.13* to *0.90.*

To summarize, the analysis of the 10 field studies included in the analysis indicates that while polygraph examinations using CQT in criminal investigations detect deceptiveness and nondeceptiveness better than chance, there is also what in some cases might be considered a high error rate, particularly for nondeceptive subjects. The one study which tested the validity of the relevant/irrelevant question technique (the general question test (GQT) portion of the Bersh study) also detected deceptiveness and nondeceptiveness better than chance.

Variation Among Studies

As implied in the introduction to this section, the use of a single statistic or summary number to describe the results of field tests of validity may be misleading. As shown in table 3, although the field studies of polygraph validity are similar in that almost all of them tested control question techniques in criminal investigations, they differ in operationalizations of ground truth and type of examiner decision. The result is that there is a great deal of variability in the results of studies. Correct guilty detections range from 70.6 percent in one condition of the Bersh study to 98.6 percent in a condition of the Wicklander and Hunter study. Correct innocent detections are even more variable, ranging from a low of 12.5 percent in the Barland and Raskin judicial outcome study to a high of 94.1 percent in one condition of the Bersh study. Table 5 also indicates the range of incorrect judgments and inconclusive among studies. False negatives range from 29.4 percent of the Bersh study to zero percent. False positives range from 75 percent in Barland and Raskin (22) to zero percent in two studies. Inconclusive range from zero to *25* percent. This section compares studies that used comparable operationalizations of construct and criterion validity in an attempt to discover reasons for the range of results. However, even using this method results in considerable variability. The main point, however, is that no field studies exist to directly test the situations for which DOD and the President propose to expand polygraph use.

Table 4.—Mean Detection Rates as a Percentage of Total in Field Studies

Examiners or evaluators' diagnosis	"Ground truth" Percent guilty Mean S. D.	Percent innocent Mean S.D.	
Deceptive	49.3 (12.7)	8.2 (7.2)	57.5
Nondeceptive.	5.8 (5.1)	32.7 (16.7)	38.5
Inconclusive	2.0 (3.0)	2.1 (2.5)	4.0
	57.1	43.0	100 %

NOTE lambda, 0.65
S D = standard deviation

Studies Using Panel Criterion and Examiners' Decisions

Both Bersh *(29)* and Barland and Raskin (22) used a panel to establish the criterion for validity in their studies. The makeup of the panels and the polygraph scoring systems were similar in each study. In the Bersh study, which validated polygraph examinations conducted by military examiners, the panel consisted of four Judge Advocate General (JAG) Attorneys; Barland and Raskin's panel consisted of two criminal defense attorneys, two criminal prosecuting attorneys, and a judge. The examiners in the Bersh study used either GQT (a version of R/I) or the zone of comparison (ZOC) technique; for all but one subject in Barland's study, the Federal ZOC control question technique was used and results evaluated using the Army scoring procedure. Assuming the accuracy of the panel's decisions, the two studies' results are strikingly different. Barland and Raskin attained accuracy rates of 91.5 percent for guilty

Table 5.—Outcomes of Field Studies of Validity

	Guilty				Innocent					
	Number of cases	Correct	Incorrect (false neaative)	Inconclusive	Number of cases	Correct	Incorrect (false Posifive)	Inconclusive	Total number of cases	Lambda
Bersh (29) (panel of 4) GOT unanimous	32	96.9%	3.1%	O[c]	36	88.9%	11.170	U[c]	W	
ZOC unanimous	38	89.5	10.5	0	51	941	5.9	0	89	
Average unanimous	70	93.2	6.8	0	87	91.5	8.5	0	157	0.84
Majority (ZOCand GQT) ,,	34	70.6	29.4	0	25	80.0	**20.0**	0	59	0.82
		81.9	18.1	0		85.6	14.3	0	216	
Horvath and Reid (84) (1 examiner, IO examiner/evaluators)	20	85.0	15.0	Ob	20	905	9.5	Ob	40	0.76
Hunter and Ash(85)–(1 examiner, 7 examiners/evaluators)	10	87.1	11.4	1.4	10	86.4	14.1	0	20	0.74
Slowick and Buckley, (155)[d]–(random selection; 7 examiner/evaluators)	15	84.0	15.3	0.7	15	90.7	6.6	2.7	30	0.77
Wicklander and Hunter (205)–(2 examiners/6 evaluators)										
PG+[e]	10	98.6	1.3	0	10	86.6	8.3	5.0	20	
PG ." ::		**90.0**	8.3	1.6		86.6	5.0	8.3		
Average		94.4	5.0	1.0		86.6	6.7	6.7		0.88
Horvath (82)[f](10 examiner/evaluators)										
Verified cases	28	77.1	22.9	09	28	51.1	48.9	09	56	0.28
Davidson (47)[h]–(random selection 7 examiners/evaluators)	10	**90.0**	10.0	0	11	91.0	0	9.0	21	0.90
Raskin,[d] [d]	12				4				16	
(1 examiner, 25'evaluators) Numerical		**91.7**	0	8.3		75.0	0	25.0		0.75
Nonnumerical		**83.3**	8.3	8.3		25.0	**50.0**	25.0		
Barland,and Raskin(22)(l examiner, panel of5)										
Panel J	47	**91.5**	0	8.5	17	29.4	52.9	17.6	64	0.29
Judicial outcome	33	**90.9**	0	9.1	8	12.5	75.0	12.5	41	0.13
Kleinmuntz and Szucko (92)–(5examiners/evaluators[k])	50	**75.0**	25.0	01	50	63.0	37.0	01	loo[b]	0.38

~Datafor mconcluswes notreporfed; **inconclusives** appear to total 27 (243 initialN—216 decisions reported)

Examiner/evaluators were not allowed to judge charts as inconclusive as to overall deceptiveness In another type of analysis on a question by question basis, Judgments of doubtful or inconclusive were allowed

[c] Average of two blind chart analyses spaced at least 3 months apart, done by same examiners

[d] Average frequencies divided by number of examiners

[e] PG+ indicates evaluators had access to writteninformation in addition to polygraph charts (e.g. case details, subject behavior during examinations, etc) Both PG only and PG+ examinations were done by the same examiners 2 months apart

[f] Excludes Horvath's analysis of 28 unverified cases, because there is no criterion, this is a reliability study

[g] There were apparently 15 (1 3 percent) inconclusive judgments out of 1,120 total judgments (10 examiners ×112 cases) which the author excluded from further analyses

[h] Majority decision only.

[i] Seven examiners used numerical scoring, 18 used nonnumerical scoring procedures

[j] Excludes 28 cases for which the panel was unable to come to a decision as to guilt or innocence

[k] Decisions were based on one polygraph chart standard practice generally employs at least three Also, the evaluations were made by students with little polygraph experience

[l] Inconclusive were not allowed

NOTE: G(IT = general question test
Zoc = zone of comparison

and 29.4 percent for innocent subjects; comparable figures in Bersh's study are 70.6-percent guilty correct and 80-percent innocent correct. It is not clear why there should be this variation, although differences in the nature of the cases, the completeness of the case files, and sample selection may account for some of the differences.

In the Bersh study, cases were initially drawn at random from a pool of criminal investigations conducted by the three military services over a period of *3* years (1963-66); then, any cases which had been judged "indeterminate" by the original polygraph examiner were eliminated. In addition, after polygraph charts were removed from the investigative files, a preliminary panel of judges eliminated from the sample all files containing insufficient evidence to warrant a positive determination of guilt or innocence. Only those cases which resulted in a unanimous decision by the initial JAG panel were retained in the validation sample. Altogether, one-quarter of the cases (80 cases out of 323) were eliminated because of insufficient evidence. This figure does not include the number initially eliminated on the basis of inconclusive polygraph examinations.

In Barland and Raskin's (22) study, the initial pool of subjects consisted of 102 (nonmilitary) criminal suspects referred to Barland by police, defense or prosecuting attorneys. These cases represented the entire population of Barland's cases at that time, Then, *92* of these 102 cases were retained for further analysis on the basis of independence (a case was considered independent where two or more subjects had not been examined regarding the same crime). In one respect (the fact that there was only one examiner), Barland and Raskin's sample was less variable than Bersh's. However, Barland and Raskin did not eliminate from consideration indeterminate examinations. Neither, and perhaps more importantly, did Barland and Raskin eliminate cases in which investigative files without the polygraph were inadequate. As Barland (17) points out, many of the investigative files that were given to the panel were incomplete. The files had been compiled by inexperienced student assistants who often did not know where to obtain necessary information. The officials responsible for providing the information were, more often than not, unavailable or, when they were available, unable to recall the details of a crime. In many cases, few details were available. As a result, one-third of the 92 cases were judged inconclusive by the panel merely on the basis of the investigative files. The figures reported in table *5* are for *64* of the original *92* cases.

It is not clear why there should be an inverse relationship between accurate detection of guilty and innocent suspects in the two studies. It may be that both the panel and the examiner in the Barland and Raskin study consistently tended to presume guilt in the absence of any a priori base rate (see 28,160). The cases in the Bersh study, on the other hand, were initially selected to be equally distributed among deceptive and nondeceptive cases. It is not reported whether the panel was aware of the base rate in the Bersh study.

Studies Using Confession as a Criterion and Blind Evaluations

The remainder of the field studies analyzed tested the validity of polygraph testing by comparing the blind evaluations of polygraph examiners against a criterion of verification by confession. Two exceptions are Barland and Raskin's judicial outcome analysis and one condition in the Wicklander and Hunter study. The confession studies vary somewhat as to source of verified files. The Horvath and Reid, Hunter and Ash, Slowick and Buckley, Wicklander and Hunter, and Kleinmuntz and Szucko studies all used files from polygraph testing firms. Horvath's cases came from police files, Davidson's from military cases, and Raskin's from the Barland cases reported in Barland and Raskin (22; discussed above). The first four studies used files from the firm of John E, Reid & Associates and involved various criminal offenses. The firm used by Kleinmuntz and Szucko is not identified; all of their cases involved theft.

In the first four studies, blind examiner evaluators also came from John E. Reid & Associates. The Reid studies did vary with respect to case selection. Only one study (Slowick and Buckley) reports random selection of cases; in other studies, the cases of only one or two examiners were used. Horvath's (82) blind evaluators were field-trained

examiners with a median of 3 years experience, all of whom specialized in conducting polygraph examinations for police agencies. The *25* evaluators in the Raskin (133) study were volunteers who had trained in a variety of places.

The results of the Reid studies do not vary substantially. The greatest deviation from the mean occurred in one condition of the Wicklander and Hunter study in which examiner/evaluators were given additional information about the suspects (verbal and nonverbal behavioral indicators, demographic information) and the **cases.** This difference, however, was not statistically significant. Even so, it maybe reasonable to consider it separately from the other Reid studies, because of the extra information available to evaluators. In the Reid studies, guilty correct identification rates ranged from *84* to 87.1 percent, with an average of 86.5 percent (excluding the 98.6-percent Wicklander result; 88.9 percent including it). The innocent correct rates in the Reid studies range from *86.4* to *90.7* percent with an average of *89* percent. There is no difference when the Wicklander and Hunter condition is included.

An additional difference of note among the Reid studies concerns the false negative rate, which is highest in the studies which either used random selection of cases (Slowick and Buckley) or eliminated the most clear-cut charts from their original selection (Horvath and Reid). There is no apparent explanation for the variation in false positive rates in the Reid studies, which ranged from 5 to **14.1** percent.

The Davidson study results are basically similar to those of the Reid studies, except for the absence of false positives. However, the study should be interpreted with caution as one-third of the originally (randomly) selected sample was not able to be used.

The Horvath *(82)* and Kleinmuntz and Szucko *(92)* studies have the lowest accuracy rates. As with the Barland and Raskin (22) study, the low accuracy rate may be related to the fact that Horvath selected his sample from police files. Perhaps, police records of verification are not reliable, or have greater variability than those of polygraph firms.

Barland (17) has suggested a number of reasons why Horvath's results are lower than the Reid studies. One reason is that the blind reviewers did not have access to "special charts" administered in *32* percent of the cases, primarily to subjects the original examiner considered deceptive; these charts were removed from the files before being reviewed by blind examiners. According to Barland, Horvath's original examiners had been 100 percent correct in their judgments. A second reason is that, as noted above, police examiners were used instead of private examiners; the difference between the two kinds of examiners is not explained further. Yet a third reason, which Barland (17) believes may be the most important in terms of false positives, is that a number of victims and witnesses were included in the sample (i. e., were subjects). According to Barland (17), one theory of detection of deception predicts that innocent victims or witnesses may react emotionally during a polygraph examination because they experienced or witnessed the event regardless of whether they are telling the truth about specific details of the incident. An analysis of the Horvath data suggested by Barland, comparing results for victims and witnesses with those for suspects, would be of interest (see Giesen and Rollison (61) for a comparison of innocent associations with guilty knowledge).

Despite the generally anomalous results of Horvath's (82) study, an interesting finding may help to account for the results of the Kleinmuntz and Szucko *(92)* study. Horvath found that suspects in crimes against property were less detectable than suspects in crimes against persons. This may be because crimes against persons are likely to have a greater amount of affect associated with them, and are, thus, more physiologically detectable. Barland and Raskin (22), on the other hand, found no differences by type of crime. As noted previously (see table 3), Kleinmuntz and Szucko's *(92)* study selected only cases from the files of a polygraph firm involving crimes of theft. However, although the crimes against property hypothesis is suggestive, it may not fully explain the difference between Kleinmuntz and Szucko's and similar studies. The Davidson study, for example, only used theft cases, and it has a "O" false

positive rate (although it has a substantial inconclusive rate). Analyses of other studies by crime type would be informative, although the number of cases would probably be too small for a meaningful analysis.

Szucko (159) has suggested that one possible reason his results are so different from other polygraph firm studies' results, is that the individual who selected the charts in the Kleinmuntz and Szucko study could not read polygraph charts. Therefore, case selection may have been more variable than in some of the other studies. Alternative explanations are that: 1) Kleinmuntz and Szucko only evaluated one chart for each subject (at least three is standard); and 2) their evaluators were examiner-trainees at the end of their internship period, not experienced examiners* (see 91).

● Some maintain that the evaluators in Kleinmuntz and Szucko's study were even less experienced than that.

Studies Using Judicial Outcomes and Original Examiners' Results

Barland and Raskin's (22) analysis using judicial outcomes as a criterion has the lowest accuracy rate for innocent suspects—a 12.5-percent innocent correct and 75-percent false positive rate. The problems with using judicial outcomes as a criterion have already been referred to, in particular, the fact that the judicial outcome is not a highly accurate measure of guilt because of such characteristics of the legal system as the necessity for proof beyond a reasonable doubt, and the prevalence of plea bargaining. These problems are illustrated here by the fact that only 41 of Barland and Raskin's original *92* cases were resolved by the criminal justice system. Again, there is clearly greater agreement on guilty subjects.

OTHER CONSIDERATIONS

Although the analysis above demonstrates that polygraph testing is better than chance at differentiating deceptive from nondeceptive subjects in criminal investigations, what might be considered as substantial false positive and false negative rates are obtained in several investigations. Although it is not possible to determine a "scientifically" acceptable rate of correct or incorrect judgments, clearly if error rates are between 10 and 25 percent, a large number of incorrect decisions would be made if the polygraph were widely employed. The base rate of guilt in actual situations may further complicate matters. It is not clear from the field studies conducted so far how many suspects were involved in the **cases** selected for polygraph testing, but if there were a large number of suspects, more false positives could be expected (see ch. *7).*

Also problematic is the wide variability in accuracy rates across studies. Although some differences can be explained methodologically, other differences cannot. Of perhaps even greater importance than the accuracy rate variability and error rate problems is the observation that field studies of polygraph testing have only been conducted in criminal investigations. As is discussed more fully in chapter *6,* criminal investigations may generate different levels of affect. In addition, different kinds of subject groups maybe the focus of expanded Government use of polygraph testing. Only two field studies can be identified that relate directly to polygraph testing in the national security area: one by the Director of Central Intelligence (DCI,165) and a second by Edel and Jacoby (51). Neither of these is a validity study but because they are the only field studies with any relevance to national security, they will be described below in some detail. An analog study of counterintelligence screening (16) is discussed in chapter 5.

The DCI study consisted of a survey of 12 Government agencies (not including the National Security Agency (NSA)). The study was conducted to evaluate the relative effectiveness of various means of conducting background investigations for purposes of applicant screening and security

clearances for current employees. Background investigations are conducted through the use of personnel interviews, interviews with present and former neighbors, checks of educational and work records, and checks with a consortium of other national agencies (the so-called National Agency Check). Of the agencies surveyed, only the Central Intelligence Agency (CIA) used the polygraph to conduct background investigations.

In the 4-month period covered by the study, CIA conducted *507* background investigations. Of these, adverse information arose concerning 47 percent of applicants or other individuals being investigated for security clearances. Thirty-five (83 percent) of the adverse cases were resolved against the individual (i. e., the applicant was not hired or clearance was not granted). In two-thirds of the instances of adverse information resolved against the individual with the use of the polygraph, subjects admitted to the adverse information. The kinds of issues admitted by subjects had primarily to do with drug and alcohol use (e.g., marijuana use, alcohol abuse, abuse of other drugs; approximately 55 percent of the cases) and immoral conduct (e. g., sexual deviance; *24* percent of cases). Four cases involved irresponsibility, a subcategory of which is violation of security regulations, and none involved the loyalty category. It is not clear whether any of the four irresponsibility cases involved violations of security regulations. Three of the eighty-four resolved against cases involved admissions of foreign connections, meaning in this case either that:

1. the subject was not a U.S. citizen;
2. the subject's spouse was not a citizen;
3. relatives were potential "hostages;"
4. alien relatives, "hostage" unlikely; or
5. life abroad cannot be verified.

The seriousness of the wrongdoings was not clear.

The crux of the DCI analysis was the construction of a productivity index for investigative techniques from the CIA data and data from other agencies. Based on the fact that a large number of cases were resolved against individuals by admission, and the polygraph was the "unique source" (165) in all the CIA cases resolved against the subject, DCI tentatively concluded that the polygraph was the most productive of all background investigation techniques. For admissions, for example, the polygraph had an index of *6.59* compared to 0.79 for "administrative screening, " 1.08 for "investigative interviews, " and 0.28 for "papers only. "

Several aspects of the study should be noted. One is that the criteria for case selection and adverse information are not stated. Another issue, noted by the DCI study authors, is that even though the polygraph is reported as the sole source in resolving adverse information, it was only used after a thorough investigation using other sources had taken place. For this reason, it is difficult to assess its effectiveness separately from the effect of a thorough investigation. Furthermore, as a result of being conducted at the end of a background investigation, in this case the polygraph examinations could be considered a confrontation technique rather than an investigative tool, according to DCI. Agencies surveyed by DCI were asked not to include confrontation techniques in their responses. A third problem is that there was no independent verification of the cases that were resolved. Perhaps most important, the effectiveness of polygraph examination cases involving most, if not all (i. e., irresponsibility) of the kinds of adverse information uncovered among applicants in the sample probably cannot be generalized to investigations of unauthorized disclosures.

Edel and Jacoby (51), in a study reported in a leading psychology journal, tested the reliability of polygrapher judgments of physiological responsivity in applicants for positions with "a large Government agency." Forty cases were randomly selected from the agency's applicants in 1966. Ten practicing polygraph examiners acted as actual examiners in four cases each and raters in eight additional cases. In each case, examiners (raters) judged three physiological responses to each interview question as either "no specific reaction" or "a specific physiological reaction. " The rate of agreement between examiners and raters as to whether a physiological reaction took place averaged 96 percent.

Of course, as the authors note, demonstrating consistency among examiners "is not equivalent

to demonstrating consistency in interpretations based on these physiological reactions" (51). For example, responses were not differentiated for relevant v. irrelevant questions. Therefore, although Edel and Jacob's study indicates that the examiners in the Government agency can reliably detect physiological reactions, whether these physiological reactions indicate deception among applicants for positions in Government agencies has not been tested. Because of the potential adverse consequences for employment applicants (particularly in Government agencies where there is interagency checking (see, e.g., *165)),* such tests have substantial practical significance.

CONCLUSIONS

Although there is some evidence from available field studies that polygraph testing is effective in detecting deception by guilty criminal suspects, there is also what in some cases might be regarded as a substantial error rate. This is particularly so for innocent subjects. There appears, as yet, to be no scientific field evidence that polygraph examinations can be effectively used to investigate unauthorized disclosures or that they represent a valid test to prescreen or periodically screen Government employees. Results of field studies are subject to additional problems of interpretation because of inadequate measures of ground truth.

The following chapter reports on the effectiveness of polygraph testing demonstrated by analog studies. As will be shown, the construct and criterion components of validity are stronger in analog studies, but because of problems with external validity, they do not provide evidence about actual polygraph testing that is as direct as that from field studies. Nevertheless, reviewing such evidence is necessary to assess both the present and potential use of polygraph testing.