ON THE FALLIBILITY OF LIE DETECTION

BENJAMIN KLEINMUNTZ
JULIAN J. SZUCKO

The polygraph's widespread use in the legal setting and elsewhere should be of concern to society, but especially to psychologists and lawyers. Since lying does not produce a measurable physiological response—and hence renders "lie detection" meaningless—the plausibility of the theory of so-called lie detection tests is questioned. Empirical evidence is presented that disputes the accuracy of testing and shows the high rate of false positive misclassification (e.g., misclassifying a truthful person as deceptive). An alternative procedure is recommended. This procedure, sometimes called the Guilty Knowledge Test, has some problems associated with its use and can be used only when particular information is available. However, it can be a significantly more accurate detector of guilt than the standard lie detection test.

The legal system depends upon its ability to judge witness credibility. Confidence in the accuracy of witness testimony may influence the police officer's decision to arrest, the prosecutor's decision to indict, the judge or jury's decision to convict. It is thus easy to understand why members of the legal community might wish for a device that would reveal deception. Such powers are claimed for the polygraph or lie detector. While the results of lie detector tests are not routinely accepted as admissible evidence, a number of states permit their introduction in some circumstances (e.g., California, Indiana, Ohio, Wisconsin).[1] Moreover, results of lie detector tests have been held as admissible according to the trial courts' discretion in several federal circuits (e.g., Sixth, Seventh).[2] The United States Supreme Court has not yet ruled upon the admissibility of polygraph test results (43 ALR Fed 68 [1979]), but in view of continuing interest in and debate about the polygraph, the Court may eventually accept a case dealing with the topic. Indeed, while there was little openness to the

[1] California: *People* v. *Houser* (1948); *People* v. *Reeder* (1976). Indiana: *Willis* v. *State* (1978); *White* v. *State* (1978). Ohio: *State* v. *Souel* (1978). Wisconsin: *State* v. *Stanislawski* (1974). (Other states include Iowa, North Dakota, North Carolina, Florida, Georgia, Kansas, and New Jersey.)

[2] Sixth Circuit: *United States* v. *Ridling* (1972). Seventh Circuit: *United States* v. *Penick* (1974).

LAW & SOCIETY REVIEW, Volume 17, Number 1 (1982)

lie detector in U.S. courts prior to 1972, the last ten years have seen some shift in position based on the presumption that techniques may now be sufficiently improved to warrant consideration. For example, in commenting on the pertinent Wisconsin case, the Federal Court of Appeals for the Seventh Circuit noted, "Indeed, the change in Wisconsin law allowing the introduction of polygraph testimony rests on the adoption of the premise in *Stanislawski* that polygraph examinations have become more reliable and have achieved such a degree of scientific recognition that their unconditional rejection is no longer appropriate" (*McMorris* v. *Israel* [1981]). In this paper we provide evidence on the validity of the device, concluding that the claims for its accuracy are inconsistent with the evidence. We then propose an alternative technique for detecting deception.

I. THE PSYCHOLOGICAL CONTEXT

Like intelligence tests such as the Stanford-Binet, or personality tests such as the Rorschach or MMPI, the lie detection procedure is a psychological test. Thus the lie detector is a psychometric device in the sense that it is a standardized instrument or systematic procedure designed to obtain a measure of a sample of psychological behavior. In this case, rather than measuring some aspect of intelligence or personality, the behavior presumably being sampled is lying, or its opposite, telling the truth. As a psychometric instrument, the polygraph examination must meet the same standards that are required of all psychological tests. Among the most important of these standards are *reliability* and *validity*. Reliability deals with the consistency with which a test measures its dimensions. This consistency or stability can be assessed by measuring the extent of agreement in scores among individuals between test and retest, or by gauging the level of consensus attained among testers who score the same on polygraph charts. Although several studies have reported rates of agreement in the 80- to 90-percent range (Hunter and Ash, 1973; Raskin *et al.*, 1978), it is difficult to determine whether this agreement indicates that the polygraph test is reliable. The problem lies in the fact that total percentage of agreement is a poor measure of reliability. If a test has perfect reliability we would expect to obtain 100 percent agreement among interpreters. However, if a test has zero reliability, the expected percentage of agreement would not be zero percent, but rather 50 percent. This is the lowest level of agreement, the

chance level, against which to compare obtained agreement rates. However, under certain conditions, the chance level of agreement may be considerably higher than 50 percent. For example, suppose that two examiners each randomly pass 90 percent of the subjects whose tests they interpret. Then if they both evaluate the same tests, we would expect them to agree by chance on 82 percent of their decisions. We would expect examiner 1 to pass 81 percent of the subjects that were passed by examiner 2 (.9x.9), and to fail one percent of the subjects that were failed by examiner 2 (.1x.1). A more appropriate measure than total percentage of agreement can be obtained by using the correlation coefficient, or by using the average percentage of agreement. Unfortunately, most studies do not report such information, and thus the reliability of the polygraph is difficult to gauge. Even if a test achieves a high reliability, however, the test may have little utility if its validity is low.

Validity evaluates the extent to which a procedure measures what it *claims* to measure. In other words, the validity of a lie test deals with the correspondence between the test results and the subject's actual behavior, or the criterion of "ground truth." Thus, while a test may be reliable in the sense of obtaining similar scores when given on different occasions, or when given by different examiners, these scores will not be valid if they are not associated with the behavior that is of interest. Therefore, to assess the validity of the polygraph examination, a relationship must be established between the polygrapher's judgments and the truthfulness of each subject's statements. As we shall show later, polygraphic interrogation does poorly indeed in this regard.

The lie detection test is concerned with psychophysiological responsivity. It purports to establish a relationship between lying and certain physiological changes. But the paradox here is that there is no reason to believe that lying produces distinctive physiological changes that characterize it and only it. In other words, there is no set of responses—physiological or otherwise—that humans omit only when lying or that they produce only when telling the truth (Lacey and Lacey, 1958). No doubt when we tell a lie many of us experience an inner turmoil, but we experience a similar turmoil when we are falsely accused of a crime, when we are anxious about having to defend ourselves against accusations, when we are questioned about sensitive topics—and, for that matter, when we are elated or otherwise emotionally stirred. In

short, as Lykken recently suggested, "the polygraph pens do no special dance when we are lying" (1981c : 10).

II. THE TECHNIQUE OF POLYGRAPH EXAMINATION

Before discussing the social impact of the lie detection test, a brief description of the polygraph procedures may be helpful. The first phase of the examination is usually the pretest interview, which is conducted without the aid of the polygraph. This is a structured interview designed to obtain biographical data and to evaluate the subject's attitudes toward dishonesty, as well as to assess his or her attitudes toward the test itself. While the primary purpose here is to develop a data base from which to formulate questions for the polygraph phase, the responses and behaviors are also treated as interpretable data (Horvath, 1973).

The second phase of the examination is generally considered the polygraph test proper. It is only during this phase that the subject's physiological reactions are monitored. The usual field examination requires the continuous recording of three to four channels of physiological data. The variables measured typically include galvanic skin response (GSR), blood pressure, abdominal respiration, and thoracic respiration. The latter two measure the amount of external stomach and chest movement by means of a system of attached tubes and bellows. Blood pressure is continuously monitored through a similar system that uses a sphygmomanometer cuff, usually attached to the bicep (for a further description of this instrumentation, see Reid and Inbau, 1977).

In the standard polygraph phase, the examiner asks a series of questions, each requiring a simple yes-no answer. These questions are formulated during the pretest interview and are reviewed with the examinee just prior to the polygraph monitoring, a review that is intended to clarify unnecessary ambiguities. As the questions are presented, both verbal and physiological responses are recorded on the polygraph chart.

The above question sequence is usually repeated three or four times, often with a stimulation test inserted between the first and second presentation. The stimulation or "card test" procedure is usually used to convince the subject of the infallibility of the polygraph. During this test, the subject is presented with several numbered cards and is instructed to select one. The examiner then tries to identify the chosen card, informing the subject that this will be done on the basis of the polygraph tracings. However, the examiner may actually use

any number of deceptions (e.g., memorizing the position of each card) to ensure that he or she correctly identifies the target card. At this point the subject may also be cautioned that it is difficult to "beat the machine."

The specific questions the subject is required to answer vary depending on the reasons for the examination, but most current examinations include three general types of questions. These have been designated as case-irrelevant, case-relevant, and control questions. Case-irrelevant questions deal with established biographic data (e.g., name, age, current address) and are designed to obtain a normal or baseline response level; case-relevant questions deal with the specific issues under investigation; and control questions attempt to force the subject to lie about some normatively shared transgression (e.g., "Did you ever steal anything in your life?"). The control questions permit the examiner to observe the subject's physiological correlates of mild emotional arousal.

In evaluating the polygraph charts, most interpreters look for signs of differential autonomic disturbance. If the disturbance associated with the relevant items seems to be greater or more persistent than that associated with the control questions, then the subject is judged to be deceptive. On the other hand, if the disturbance associated with the control questions appears to be greater or equal to that of the relevant question, then it is assumed that the subject is being truthful. Most interpreters use a global evaluation method without specific measurement or scoring, although a method which requires that the autonomic differences between relevant and control items be assigned numerical values or scores has been introduced (Backster, 1963; Raskin and Hare, 1978). These scores can then be used in evaluating the subject's responses. In the experiments reported below, we gave to the judges as evidence the physiological data from a single question sequence, not the entire examination protocol, and we compared the clinical processing of these data by judges with that achieved by statistical formulas and methods in response to the same data.

III. THE LEGAL AND SOCIETAL CONTEXT

There are many legal settings in which polygraph examinations play an important role. These settings include those in which the test is given to suspects in criminal investigations who have been accused of theft, rape, murder, or of lesser crimes (Cimmerman, 1981); to victims of crimes who

may be so unfortunate as to reside in counties or states where the district attorney will not prosecute a complaint unless the victim agrees to a lie detector test; or to witnesses in a criminal trial who are requested by either the prosecuting or defending counsel to demonstrate the veracity of their testimony (Lykken, 1981c). Other legal contexts include those in which polygraph examinations are commonly given to plaintiffs or defendants embroiled in civil litigation in which conflicting testimony must be weighed. In most of these instances, the polygraph procedure, although not necessarily admissible as evidence in a court of law, nonetheless can contribute in important ways to the conduct and the outcome of litigation. For example, the senior author recently appeared as an expert witness for the defense in the case of a policeman who was accused of burglary and subsequently gave a classic "deceptive" pattern on tests administered on two separate occasions by the same polygraph firm. Up to the time of the polygraph investigations, the prosecution had a weak case and was prepared to drop charges if the defendant agreed to submit himself to a polygraph examination, and if he "passed" the test. The officer's defense attorney firmly believed in the innocence of his client, and had faith in the test's ability to exonerate the police officer. He therefore counseled his client to accept the prosecution's offer. He also admonished the defendant of the consequences of "failing" the test—namely, the prosecution's case would be strengthened. Although the police officer was quite apprehensive about the consequences of "failing" the test, and somewhat concerned that a refusal to comply with the prosecution's request might be interpreted as evasiveness or even guilt, he was convinced of his innocence and therefore agreed to submit to a lie detection procedure. There was a great deal at stake for him. He was accused of a felony, was already suspended without pay, and was upset about the possibility of further disgrace, perhaps even imprisonment. All these factors made him a less than ideal candidate for polygraphy.

Predictably, the officer produced polygraph records on both testing occasions that were consistent with "deception." In other words, his responses were more physiologically reactive to the following three relevant questions than to a set of control questions that dealt with some minor transgressions in his youth:

1. "Did you case Mr. and Mrs. X's house on the night of July 15th?"

2. "Did you steal the missing items from Mr. and Mrs. X's house on the night of July 15th?"
3. "Did you break into the rear door and enter the home of Mr. and Mrs. X on the night of July 15th?"

It should be mentioned that the home was indeed broken into on the designated night and that it was reported by the policeman under investigation the morning after the break-in. In fact, as it turned out, the police officer had been asked by Mr. and Mrs. X to "look after" the house during his off-duty hours, an arrangement which was against department regulations. The officer's strong polygraphic reaction may very well have been caused by his realization of the illegality of his moonlighting activities. Thus although he was not casing the house, he was nonetheless in violation of a rule, but one whose violation was not a felony and did not carry a prison term. He was clearly upset. He became even more upset over the next question—and showed an even stronger emotional reaction, because it became obvious to him that his case was becoming weaker with each question asked. Of course his emotional reaction to the third question was unmistakably greater than it had been to any of the control questions (e.g., "Did you ever steal anything of value?") because he was sensible enough to understand that his emotional responses to this question could jeopardize his future and his reputation.[3]

But the impact of lie detection testing extends beyond the legal system. According to one source (Lykken, 1981a), in 1980 alone one million Americans were subjected to lie tests. These tests were used for screening job candidates applying for sensitive positions in federal security agencies (FBI, CIA, NSA) and for many state and local police department jobs, as well as bank and armored truck jobs, all of which require honesty and integrity. Even in less sensitive positions, lie detecting tests are often given in an effort to predict which employees can best be trusted to handle large sums of money or other valuables. A recent survey indicated that 20 percent of the nation's major corporations use polygraph testing of employees, and about 50 percent of fast-food companies like McDonald's 4,700 hamburger outlets and Burger Chef's 900 stores use pre-employment polygraphs for screening (Belt and Holden, 1978). These government agencies and retail outlets

[3] The judge in this case was persuaded by expert testimony that no unique lie response will show up when a person is lying that will not register when that person is upset. He was also receptive to evidence that polygraph interpreters, regardless of their many years of experience, often have only 60-70 percent accuracy, with false positive rates in the 30-50 percent range.

are mainly interested in identifying and hiring persons who can be trusted to fulfill their jobs faithfully and honestly. Unfortunately, the evidence suggests that these enterprises, as well as those that subject their employees to "periodic honesty checks," may be wasting their time and money, and may be doing a disservice to many job applicants and employees who are turned away or fired on the basis of these lie detection devices.

As we shall show next, no evidence has been presented so far—either in the psychological or polygraphic literature—that demonstrates a high correlation between polygraph results and the criterion or "ground truth" of honesty.

IV. THE FALLIBILITY OF LIE DETECTION

Having established that the lie detector is a psychological device that is widely accepted and commonly applied in a large variety of settings, we can now ask: how fallible is the polygraph examination? This is a question that has stirred lively debate in the lie detection literature between those proponents who advocate its use as a valid technique to discriminate between truth and deception, and the skeptics who do not. The first are represented by Podlesny and Raskin (1977) who report that the false positive rate (i.e., proportion of persons incorrectly identified as lying) is within the range of two percent to eight percent, and who conclude that a number of physiological measures are accurate indices for discriminating between truth and deception. This position is seriously challenged by Lykken (1979), who asserts that the false positive polygraph expectancies are much higher, more likely within the range of 36 percent to 39 percent, and who argues that the theory of the lie test is implausible because it purports to identify a psychological state that has no characteristic physiological index. On this basis Lykken concludes that one should not accept the claims of very high accuracy that are sometimes made by the technique's advocates (see also the rejoinder by Raskin and Podlesny, 1979).

Most empirical studies of the test's validity can be classified into two general categories. The most common strategy, using the physiological detection of deception paradigm, has attended to the validity of the technique's physiological measures in predicting (or detecting) lying (Davidson, 1968; Gustafson and Orne, 1963; 1964; 1965; Kugelmass *et al.*, 1967; Lykken, 1959; 1960). Such studies

usually require subjects to select a numbered card, or some other information, from a set of clearly defined alternatives. The experimenter then tries to identify the information which the subject has been instructed to conceal. The physiological measures employed usually include only a single modality, most frequently the GSR, and decisions are made on the basis of a well-defined criterion. For example, when GSR has been employed, maximum amplitude has been used as the index of deception. While some of these studies have demonstrated that physiological variables can be used to detect deception, or more accurately, attempts to conceal information, they cannot be used as evidence for the validity of the standard polygraph technique. These studies have used more objective assessment procedures, thereby eliminating or minimizing the effects of the interpreter who must combine, analyze, and synthesize polygraph protocols prior to arriving at "deceptive" versus "nondeceptive" decisions. The second category consists of studies designed to evaluate the accuracy of interpreters using the polygraph technique. These have been conducted primarily by polygraph professionals (Hunter and Ash, 1973; Slowik and Buckley, 1975; Horvath and Reid, 1971; Wicklander and Hunter, 1975). Although the results from some of these studies seem to indicate that the polygraph examination is highly effective for detecting deception, confidence in these findings is mitigated by serious methodological flaws. Horvath and Reid (1971) for example, discarded 47 percent of the cases originally selected for their study on the grounds that "the evidence of truth or deception would have been obvious to any trained polygraph examiner." Furthermore, the criterion against which the interpreter's decisions were validated has not always been clearly defined. Many of these reports do not indicate whether verification of the original decisions was obtained through confessions, convictions, independent investigations, or other methods. And finally, since these have been in-house studies, conducted by firms specializing in polygraph examinations, we do not know how many of the same cases or the same interpreters were used in more than one study.

To overcome such problems, a study was conducted by the present authors (Szucko and Kleinmuntz, 1981) which focused both on the ability of the human judge as the prime decision maker, and on the extent to which polygraph tracings contain potentially useful information for discriminating between truth

and deception.[4] Thirty undergraduate volunteers participated in a polygraph experiment using a mock theft paradigm in which the guilty subjects were required to participate in the theft of some items from an office. Fifteen subjects were assigned to the theft condition and 15 were assigned to the no-theft condition. Those in the former condition were given instructions for locating an office, in which they were to search through a desk and then steal anything they desired. The desk contained a number of undesirable objects plus a five-dollar bill. All theft condition subjects elected to steal the five-dollar bill. The no-theft subjects were instructed to take a brief walk around campus prior to returning to the experimenter's office. After completing their tasks all subjects were administered a standard control question polygraph examination to determine their guilt or innocence. The polygraph tracings from these examinations were then submitted for interpretation to six experienced polygraph examiners. The examiners knew that half of the subjects had stolen something, but did not know what was stolen by whom, or that the polygraph records were obtained from experimental subjects. The results showed that the human polygraph examiners were highly fallible and that they performed less accurately than a discriminant function analysis of the polygraph data. In other words, the statistical analyses of this study demonstrated that although the polygraph charts contained some of the information necessary for discriminating between deceptive and nondeceptive respondents, the clinical judges, the polygraph examiners, were not able to make this differentiation at much beyond the chance level. The prior experience of the six polygraphers in this study ranged from 3 months to 8 years, and their correlations of truthful-untruthful decisions with the truthful-untruthful criterion ranged from .02 to .43, against the discriminant function correlation of .52. Moreover, there was no evidence that greater experience contributed to greater accuracy.

Specifically, the three polygraphers who had 3 months of experience had correlations with the criterion of .02, .23, and .43; those with 1, 2, and 8 years experience had correlations with truthful-untruthful criterion of .27, .33, and .08 respectively. Thus in both groups of judges there was at least one judge

[4] This research was conducted within the framework of the Brunswikian lens model (Brunswik, 1952), which has been used extensively to investigate human judgment (Hammond and Summers, 1972) and has been described as a more complete research paradigm than that typically used to study the judgment process (Petrinovich, 1979).

whose decisions were totally unrelated (.02, .08) to the task under investigation, and at least one polygrapher whose performance was significantly related (.43, .33) to the task criterion. These results, particularly at the low end of the performance scale, are not very encouraging, especially when we consider the evidence of some researchers (Goldberg, 1970; Hammond, Summers and Deane, 1973) who have shown that it is exceedingly difficult to improve human decision making of this sort even when decision makers are given extensive feedback about their errors.

It has been argued (see Lykken, 1981a) that a meaningful study of lie test accuracy must be done in the field with real criminal suspects; laboratory studies such as the one just described that use volunteer subjects and mock crimes do not carry the same stakes and may not produce the same emotional reaction as would real perpetrators and crimes. But conducting studies in the field is difficult for a number of reasons. Although many polygraphers are willing to cooperate in any effort that may improve the accuracy of their equipment as well as their interpretations, they are reluctant to have outsiders (i.e., psychologists) obtain access to their records. Since they have a strongly vested interest in demonstrating the reliability and validity of lie detection, they are less than enthusiastic about subjecting their methods to a test which may show that their overall accuracies are not much better than chance, and that they commit many false positive errors. These concerns, plus an understandable unwillingness to violate the confidentiality of their clients' records, have meant that the few field studies of lie detectors have generally been conducted by polygraphers themselves.

In an effort to both study and potentially improve lie detection, we were fortunate to be able to conduct a field study (Kleinmuntz and Szucko, 1982) in which we as outsiders obtained access to records in a leading polygraph firm. We selected the polygraph charts of 50 innocent subjects and 50 guilty subjects, all of whom were tested in theft-related investigations. Guilty subjects were considered verified if they confessed to the full amount of the theft of which they were suspected. Individuals were considered to be innocent if others had confessed to the offense they had been accused of. The polygraph charts of these 100 subjects were then submitted for interpretation to six polygraph trainees at the end of their internship training period. These judges were asked to rate each subject's chart on an eight-point rating scale, with a rating

of 1 labeled "Definitely Truthful" and a rating of 8 labeled "Definitely Untruthful." The interpreters were informed that all tests were obtained from theft-related investigations, and that 50 percent of the subjects were guilty. When decisions were based on only a single-question sequence, the interscorer reliabilities, the correlations between the interpreters' decisions, ranged from a low of .24 to a high of .56. The correlations between the truthful-untruthful criterion on the one hand, and the judges' ratings on the other, ranged from a low of .26 to a high of .52. A multiple regression analysis of the physiological data produced a correlation of .53. When the polygraph interpreters were allowed to use the complete polygraph chart, including all repetitions of the question sequence, the reliabilities ranged from .45 to .61, while the validities ranged from .45 to .55. Again, as in our laboratory study conducted in the same firm, we found that clinicians do not use the available polygraph data optimally and do not apply their decision rules consistently. Such nonoptimality and internal inconsistency add error variance to the clinicians' interpretations and, in turn, contribute to the inaccuracy of polygraph interpretation.

A crucial finding in this study was that on the average, the polygraph interpreters misclassified 37 percent of the innocent subjects as guilty. This high false positive rate (i.e., calling an innocent person guilty) includes, from among six judges, one who misclassified as many as 50 percent of innocent suspects, and another who misclassified only 18 percent of his innocent cases as guilty. Similar findings were found in a field study conducted by Horvath (1977), whose polygraphers called nearly half (49 percent) of the innocent and truthful respondents guilty and untruthful. Barland and Raskin (1976), in an unpublished paper, similarly report that more than half (55 percent) of their subjects were misclassified as deceptive. These data provide strong evidence that the lie detector test is indeed highly fallible and that this fallibility translates into a strong bias against the innocent respondent.

V. IMPLICATIONS OF THE FALLIBILITY OF LIE DETECTION

The meaning of this strong bias against innocent subjects can be appreciated more fully if we consider the effects of the *base rate* on accuracy. The base rate of a condition (such as paraplegia, schizophrenia, or being guilty of a certain crime) refers to the positive instances (or frequency) of that condition

among the members of a specified population. The lower the base rate, the more difficult it is, other things being equal, to identify individuals in the small sub-group and at the same time avoid misclassifying others as sub-group members (Meehl and Rosen, 1955). What is the base rate of lying among the defendants whose stipulated polygraph tests are admitted into evidence in American courts? Lykken (1981b) has suggested that prosecutors offer a defendant the option of submitting to a lie test only when the prosecution's case is weak and unlikely to sustain a conviction. If this is correct, then the base rate for lying in this population may be low, perhaps 20 percent. Under these circumstances, we are not dealing with criminal defendants in general—for whom the base rate of lying may be much higher—but rather with that subset of defendants against whom persuasive evidence of guilt is not available. In other words, among this select subset, one possible reason that the prosecution's case is weak may be that the defendant is in fact innocent. For purposes of illustration, we shall assume that 80 percent of a particular subset of 100 defendants is innocent, which means that the base rate for lying among this group will be 20 percent. Then, with a false positive rate of about 40 percent, which seems to be par for most laboratory and field studies, 40 percent of the truthful subjects, or 32 persons, will fail the polygraph test. In other words, in order to detect the 20 persons who are truly guilty, an additional 32 will be inaccurately classified as guilty. In 62 percent of the 52 cases labeled guilty, the decision would be in error.

Aside from these statistical considerations, an additional and more subtle biasing factor must be understood. This factor is a motivational one: polygraphers are motivated to serve their paying clients. Since clients have an interest in identifying guilty suspects, the polygraphers must expect to uncover cases of deception. Governmental agencies and corporations that want to ferret out security risks or dishonest persons in their organizations expect the polygrapher to identify such people. These firms often also retain polygraphers for employment screening and "periodic honesty checks," a lucrative source of income for many polygraphers.

The source of the problem is that such agencies may be more troubled by false negative errors—that is, errors that occur because a polygrapher classifies a guilty or deceptive employee as innocent or truthful—than by false positive errors. From the vantage point of a bank or security agency, it is far better to err on the side of caution and perhaps even fire (or

not hire) a trustworthy person than to run the risk of retaining a potential thief. Parole boards and mental hospitals housing assaultative inmates share a similar reluctance to bet on the low dangerousness of the individual they must judge and release. The high false positive misclassification rate avoids the potential blame a releasing agency may receive for a false negative error.

This motivation is unmistakably communicated to polygraph firms in the form of client loyalty and referrals. But it seriously compromises the polygrapher's objectivity and biases the findings against the nonpaying client, who is likely to be an individual with limited resources and is unlikely to have the power of a repeat player (Galanter, 1974). While the polygrapher is unlikely to receive feedback about his false positive errors, he is quite likely to learn about false negatives in the form of client complaints. Yet this inflated misclassification may direct law enforcement efforts in the wrong direction and can cost organizations the loss of 30 percent or more of their actual or potential work force in order to detect a much smaller number of risky candidates or employees.

VI. AN ALTERNATIVE TO THE DETECTION OF DECEPTION

An alternative and more promising method for detecting deception has been proposed, originally by Hugo Munsterberg (see Lykken, 1981a: 249), and more recently by David Lykken (1959; 1960; 1974; 1981a). This method, called the Guilty Knowledge Test (GKT), requires that the examiner be able to determine a number of facts that only a guilty subject can recognize. These facts are then presented to the respondent in the form of multiple choice items, embedded in a set of three, four or five alternatives that would seem equally plausible to an innocent subject without guilty knowledge. For example, if the crime under investigation were a jewelry store robbery in which a diamond ring was taken, then the question presented to the suspect might be "Did you steal a _____?" with watch, necklace, ring, bracelet, and brooch included as the alternatives.

The basic assumption of GKT, according to Lykken and some others who have used it, is that the guilty subject will show greater autonomic responsivity to what he or she recognizes as the significant alternative than would a subject without such guilty knowledge. The amplitude of the autonomic responses to the significant alternative has little

meaning by itself; a hyperreactive subject might respond strongly to that alternative without knowing that it was the "correct" one, while a hyporeactive person might yield a weak response even though he or she had guilty knowledge. But the same subject's responses to the other plausible but incorrect alternatives of the guilty knowledge test provide a nearly ideal control against which to evaluate his or her response to the significant alternative.

Is the GKT a reasonable alternative to current practices of lie detection? The empirical evidence seems to suggest that it may well be, although none of the studies to date have used actual criminal suspects as subjects. One early study by Lykken (1959) used student volunteers who enacted mock crimes of "theft," "murder," or both, and who were interrogated while having their electrodermal responses (EDR) recorded. The EDR is a wavelike change in the electrical resistance of the palms and soles associated with imperceptible sweating in those regions. Lykken presented the subjects with six multiple-choice questions for each of the two mock crimes. Some questions related directly to the criminal act, such as the location in which the stolen item was hidden, while others dealt with incidental matters, such as the presence of an artist's easel in the murder room. The test for each crime was scored by awarding two points if the EDR produced by the correct or "relevant" alternative to a question with the largest, one point if it was the second largest of the five associated with each question, and zero points otherwise. Thus, with a six-item test, the highest guilt score would be 12, and the most innocent-appearing score would be zero. A suspect was classified innocent if the total score for a test was 6 or less. The results were that of the 48 innocent subjects who were tested, all were identified as such; of the 50 guilty suspects tested, 44 received scores that permitted them to be correctly classified. Thus the false positive error rate was .00, while the false negative error rate was .12—an average accuracy rate of 94 percent.

In a second study (Lykken, 1960), 20 subjects were offered a money prize if they could "beat" the test, either by inhibiting their responses to the correct alternatives (which is difficult) or by producing artificial response to nonrelevant alternatives through self-stimulation (which is easier). These subjects were attached to the EDR apparatus and given time before the test to practice whatever technique they had decided to use, watching the pen trace out their EDRs as they experimented. All 20 subjects were correctly classified by the GKT, even

though they had these opportunities to practice "beating" the machine.

These studies of the GKT have been successfully replicated in numerous laboratories (Davidson, 1968; Ben Shakkar *et al.*, 1970; Lieblich *et al.*, 1976; Waid *et al.*, 1978), but not always with the high correct classification rate attained by Lykken. However, a virtue of the GKT method—one not shared by the more conventional procedures—is that the discrimination of guilty from innocent suspects can be increased simply by increasing the number of items. With 10 items, for example, each having 5 alternatives, one might expect to identify 97 percent of guilty suspects if persons scoring 5 or higher are classified as guilty. The odds against an innocent person scoring so high would be more than 100 to 1. According to Lykken (1981a: 298), this suggests that the experiments that failed to achieve good discrimination either had too few items or else used items in which the guilty subject could not recognize the "correct" alternative, or in which even the innocent subject could recognize this alternative. Lykken (1981a: 300) has developed an elaborate scheme for predicting precisely how accurate the GKT would be for tests of increasing length.

Two laboratory studies in fact reported results that fit reasonably well with Lykken's predictions. Geisen and Rollison (1980) required 20 guilty subjects to enact a mock crime and then tested them, along with 20 innocent suspects, using six well-constructed GKT items. The results were 100 percent and 95 percent, respectively, of innocent subjects who "passed" and guilty subjects who "failed." Podlesny and Raskin (1978) used five reasonably good items according to Lykken's criteria and correctly identified 100 percent and 90 percent of the innocent and guilty experimental suspect subjects.

In spite of these promising indications, professional polygraphers still have not taken up the Guilty Knowledge Test for use in criminal investigation. One reason for this is that the examiner may not have available to him or her the necessary items of information that only a guilty suspect would recognize and that could be translated into GKT items. But there are many instances of crimes which could provide the test constructor with information that would permit him or her to formulate good items. Given the reality of criminal investigation, particularly investigations of homicide cases, however, close cooperation is essential between the

investigating authorities and the polygrapher. Such cooperation is necessary because it is important that the particulars of a crime be known only to the investigators, the polygrapher, and the truly guilty suspect.

In addition to the greater effort required for test construction, polygraphers prefer conventional lie detection to the GKT, because the latter may not be an appropriate technique for about 90 percent of the situations in which lie detectors are typically used. Business, as we indicated earlier, comes mainly from "periodic honesty" checks and from tests that are supposed to predict honesty and integrity in future (or present) employment situations. The GKT procedure is totally inappropriate for these purposes, since it is only usable with reference to a specific past event.

VII. SUMMARY AND CONCLUSIONS

What can we conclude about the degree of fallibility of the polygraph examination as it is practiced today? First, our brief survey has shown that there are many legal and other societal contexts in which lie detection plays an important role. Second, we have seen that lie detection is based on psychological principles of testing and deals with psychophysiological responsivity. The polygraph examination is based on the faulty premise that suspects will experience more arousal to case-relevant questions when they are guilty than when they are innocent, and that the measurable physiological responses accompanying these arousals betray lying.

We have presented evidence showing that polygraph judges have a high rate of misclassification and that the particularly damaging by-products of these errors are false positive judgments which may label as many as 50 percent of innocent suspects as guilty. We have also argued that there are motivational factors that bias polygraphers in a way that causes false positive errors.

It should be noted that our studies, like those of other researchers, have dealt only with the issue of physiological data interpretation; the judges' decisions were based on the polygraph tracings alone. Ordinarily, interpreters would integrate such data with behavioral observations and other information collected during the testing session. Therefore, although we have dealt with the most important aspect of the test, what needs to be especially established is whether similar results would be obtained when judges are given the

opportunity to interpret complete protocols which include biographical data, case facts, and direct observations of behavior, as well as the polygraph charts. While it is possible that such data would improve accuracy, information on human decision making suggests that there would be little improvement in validity and that reliability might actually decrease (Sines, 1959).

Finally, we suggest an alternative and perhaps more promising method of lie detection called the Guilty Knowledge Test, which may be applicable in many legal contexts. At the present time, the empirical evidence that has been brought to bear on this technique—all dealing with mock experiments and volunteer suspects—suggests that the GKT is a promising method of inquiry. However, the GKT, which requires a good deal of preparation and planning, has not thus far enjoyed the popularity (or notoriety) of its more fallible counterpart. As the GKT does not readily lend itself to an employment setting, the testing situation which supplies about 90 percent of the polygrapher's routine lie detection caseload, it is is unlikely that the GKT will receive as much attention as the typical lie detection test—unless or until the low validity of available lie detection procedures is generally acknowledged.

REFERENCES

BACKSTER, Clive (1963) "Total Chart Minutes Concept," 11 *Law and Order* 77.

BARLAND, Gordon and David C. RASKIN (1976) "Validity and Reliability of Polygraph Examinations of Criminal Suspects." Report No. 76-1, Contract 75-NI-99-0001 U.S. Department of Justice.

BELT, J. and P. HOLDEN (1978) "Polygraph Usage among Major U.S. Corporations," 57 *Personnel Journal* 80.

BEN SHAKHAR, Gerson, Israel LIEBLICH, and Sol KUGELMASS (1970) "Guilty Knowledge Technique: Application of Signal Detection Measures," 54 *Journal of Applied Psychology* 409.

BRUNSWIK, Egon (1952) *The Conceptual Framework of Psychology*. Chicago: University of Chicago Press.

CAVOUKIAN, Ann and Ronald J. HESLEGRAVE (1980) "The Admissibility of Polygraph Evidence in Court: Some Empirical Findings," 4 *Law and Human Behavior* 117.

CIMMERMAN, Adrian (1981) "The Fay Case," 8 *Criminal Defense* 7.

DAVIDSON, Phillip O. (1968) "Validity of the Guilty Knowledge Technique: The Effects of Motivation," 52 *Journal of Applied Psychology* 62.

GALANTER, Marc (1974) "Why the 'Haves' Come out Ahead: Speculations on the Limits of Legal Change," 9 *Law & Society Review* 95.

GEISEN, Martin and Michael ROLLISON (1980) "Guilty Knowledge Versus Innocent Associations: Effects of Trait Anxiety and Stimulus Context on Skin Conductance," 14 *Journal of Research in Personality* 1.

GOLDBERG, Lewis R. (1970) "Man vs. Model of Man: A Rationale Plus Some Evidence, for a Method of Improving on Clinical Inferences," 73 *Psychological Bulletin* 422.

GUSTAFSON, Lawrence and Martin T. ORNE (1963) "Effects of Heightened Motivation on the Detection of Deception," 47 *Journal of Applied Psychology* 408.

——— (1964) "The Effects of Task and Method of Stimulus Presentation on the Detection of Deception," 48 *Journal of Applied Psychology* 383.

——— (1965) "The Effects of Verbal Response on Laboratory Detection of Deception," 48 *Psychophysiology* 10.

HAMMOND, Kenneth R. and David A. SUMMERS (1972) "Cognitive Control," 79 *Psychological Bulletin* 58.

HAMMOND, Kenneth R., David A. SUMMERS, and Donald H. DEANE (1973) "Negative Effects of Outcome Feedback on Multiple-Cue Probability Learning," 9 *Organizational Behavior and Human Performance* 30.

HORVATH, Frank (1973) "Verbal and Nonverbal Clues to Truth and Deception during Polygraph Examinations," 1 *Journal of Police Science and Administration* 138.

——— (1977) "The Effects of Selected Variables on Interpretation of Polygraph Records," 62 *Journal of Applied Psychology* 127.

HORVATH, Frank and John E. REID (1971) "The Reliability of Polygraph Examiner Diagnosis of Truth and Deception," 62 *Journal of Criminal Law, Criminology and Police Science* 276.

HUNTER, Fred L. and Philip ASH (1973) "The Accuracy and Consistency of Polygraph Examiners' Diagnoses," 1 *Journal of Police Science and Administration* 370.

KLEINMUNTZ, Benjamin and Julian J. SZUCKO (1982) "Is the Lie Detector Valid?" 9 *Criminal Defense* 12.

KUGELMASS, Sol, Israel LIEBLICH and Zeev BERGMAN (1967) "The Role of 'Lying' in Psychophysiological Detection," 3 *Psychophysiology* 312.

LACEY, John I. and Beatrice C. LACEY (1958) "Verification and Extension of the Principle of Autonomic Response-Stereotype," 71 *American Journal of Psychology* 50.

LIEBLICH, Israel, Gerson BEN SHAKHAR and Sol KUGELMASS (1976) "Validity of the Guilty Knowledge Technique in a Prisoners Sample," 61 *Journal of Applied Psychology* 89.

LYKKEN, David T. (1959) "The GSR in the Detection of Guilt," 43 *Journal of Applied Psychology* 385.

——— (1960) "The Validity of the Guilty Knowledge Technique: The Effects of Faking," 44 *Journal of Applied Psychology* 258.

——— (1974) "Psychology and the Lie Detector Industry," 29 *American Psychologist* 725.

——— (1979) "The Detection of Deception," 86 *Psychological Bulletin* 47.

——— (1981a) *Tremor in the Blood: Uses and Abuses of the Lie Detector*. New York: McGraw-Hill.

——— (1981b) "Letter: To Tell the Truth, " *Discover* (February) 10.

——— (1981c) "The Lie Detector and the Law," 8 *Criminal Defense* 19.

MEEHL, Paul E. and Albert ROSEN (1955) "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores," 52 *Psychological Bulletin* 194.

PETRINOVICH, Lewis (1980) "Probabilistic Functionalism: A Conception of Research Methods," 34 *American Psychologist* 373.

PODLESNY, John A., and David C. RASKIN (1977) "Physiological Measures and the Detection of Deception," 84 *Psychological Bulletin* 782.

——— (1978) "Effectiveness of Techniques and Physiological Measures in the Detection of Deception," 15 *Psychophysiology* 344.

RASKIN, David C., Gordon H. BARLAND, and John A. PODLESNY (1978) "Validity and Reliability of Detection of Deception." National Institute of Law Enforcement and Criminal Justice.

RASKIN, David C. and Robert D. HARE (1978) "Psychopathy and Detection of Deception in a Prison Population," 15 *Psychophysiology* 126.

RASKIN, David C. and John A. PODLESNY (1979) "Truth and Deception: A Reply to Lykken," 86 *Psychological Bulletin* 54.

REID, John E. and Fred E. INBAU (1977) *Truth and Deception: The Polygraph ("Lie Detection") Technique*. Baltimore: Williams and Wilkins.

SINES, Lloyd K. (1959) "The Relative Contribution of Four Kinds of Data to Accuracy in Personality Assessment," 23 *Journal of Consulting Psychology* 483.

SLOWIK, Stanley M. and Joseph P. BUCKLEY (1975) "Relative Accuracy of Polygraph Examiner Diagnoses from Respiration, Blood Pressure, and GSR Recordings," 3 *Journal of Police Science and Administration* 303.

SZUCKO, Julian J. (1982) "A Field of Study of Human and Statistical Detection of Deception." Submitted to *Science*.
SZUCKO, Julian J. and Benjamin KLEINMUNTZ (1981) "Statistical Versus Clinical Lie Detection," 36 *American Psychologist* 488.
WAID, William M., Emily Carota ORNE, Mary R. COOK, and Martin T. ORNE (1978) "Effects of Attention, as Indexed by Subsequent Memory, on Electrodermal Detection of Information," 63 *Journal of Applied Psychology* 728.
WICKLANDER, Douglas E. and Fred L. Hunter (1975) "The Influence of Auxiliary Sources of Information in Polygraph Diagnoses," 3 *Journal of Police Science and Administration* 405.

CASES CITED

McMorris v. *Israel*, 643 F.2d 458, 1981.
People v. *Houser*, 85 Cal App 2d 686, 193 P2d 937, 1948.
People v. *Reeder*, 65 Cal App 3d 235, 135 Cal Rptr 421, 1976.
State v. *Souel*, 53 Ohio St. 2d 123, 372 NE2d 1318, 1978.
State v. *Stanislawski*, 62 Wis 2d 730, 216 NW2d 8, 1974.
United States v. *Penick*, CA 7 III, 496 F2d 1105, cert den 419 U.S. 897, 42 L Ed 2d 141, 95 S Ct 177, 1974.
United States v. *Ridling*, DC Mich, 350 F Supp 90, 1972.
White v. *State*, 381 NE2d 481, 1978.
Willis v. *State*, 374 NE2d 520, 1978.