Journal of Applied Psychology
1992, Vol. 77, No. 5, 757-767

Copyright 1992 by the American Psychological Association, Inc.
0021-9010/92/$3.00

Detection Measures in Real-Life Criminal Guilty Knowledge Tests

Eitan Elaad, Avital Ginton, and Noam Jungman
Division of Identification and Forensic Science
Israel National Police Headquarters
Jerusalem, Israel

The present study provides a first attempt to compare the validity of the respiration line length (RLL) and skin resistance response (SRR) amplitude in real-life criminal guilty knowledge tests (GKTs). GKT records of 40 innocent and 40 guilty Ss, for whom actual truth was established by confession, were assessed for their accuracy. When a predefined decision rule was used and inconclusive decisions were excluded, 97.4% of the innocent Ss and 53.3% of the guilty Ss were correctly classified with the SRR measure. For the RLL measure, the respective results were 97.2% and 53.1%. The combination of both measures improved detection of guilty Ss to 75.8% and decreased detection of innocent Ss to 94.1%. The combined measure seems to be a more useful means of identifying guilty suspects than each physiological measure alone. The results elaborate and extend those obtained in a previous field study conducted by Elaad (1990).

Several polygraph interrogation techniques are used in field practice (Saxe, Dougherty, & Cross, 1985). The most common one is the controversial control question technique (CQT; Reid & Inbau, 1977; for a more detailed description of the controversy see Lykken, 1974, 1988a; Raskin, 1988). An alternative, less controversial technique is the guilty knowledge test (GKT), which is used widely in laboratory experiments but much less in field practice. The GKT consists of several multiple-choice items. One of those items is related to the crime and is assumed to be known only to people familiar with the crime. The other items are unrelated to the crime but are equivalent to the relevant item in all other respects. For example, a typical question in a murder case might relate to the way the victim was murdered. The victim could have been stabbed, strangled, shot, struck, and so on. The GKT question could be formulated in the following way: "Do you know that the victim was ____?" Then a set of alternatives is presented: stabbed, strangled, shot, struck. The crime-related item (or the relevant item) has a special meaning only for the guilty suspects, and therefore only those suspects are expected to show differential responsivity to it. A comparison of the responses to the relevant item with the responses to the irrelevant alternatives should indicate whether a subject is familiar with the guilty knowledge. If a subject's physiological responses are consistently greater to the relevant item than to the irrelevant ones, knowledge about the crime is inferred. The assumption is that subjects who have knowledge about a crime are involved in that crime unless a reasonable explanation for this guilty knowledge is provided (for example, the information was published in a newspaper). Innocent subjects, who are unable to distinguish relevant from irrelevant alternatives, are not expected to respond differentially to the relevant and irrelevant items.

Several studies, conducted in laboratory settings with electrodermal response amplitude as the only autonomic measure, have indicated that the GKT is a highly valid method for differentiating between guilty and innocent subjects (Davidson, 1968; Giesen & Rollison, 1980; Lykken, 1959; 1960; Podlesny & Raskin, 1978). Recently, Ben-Shakhar and Furedy (1990) reviewed the accuracy rates obtained in 10 GKT studies and reported mean accuracy rates of 84% and 94% among guilty and innocent subjects, respectively. These results indicate that false-negative errors are more likely than false-positive ones in a GKT examination. There are, however, some substantial differences between laboratory settings and real-life criminal situations. For example, the motivation of the subject to appear innocent should be stronger in the field situation than in the laboratory, where the guilty subjects usually do not have much at stake, and where they are aware that whatever the result of the test might be, they will be released from further involvement with the "crime." Therefore, only adequate field research can determine if the GKT is as accurate in actual criminal investigations as in the laboratory. A first attempt to evaluate the GKT in real-life criminal investigations was made recently by Elaad (1990). This study, in which skin resistance response (SRR) amplitude was used as the single autonomic measure, revealed that a very high accuracy rate (97.9%) was obtained for innocent examinees, similar to that expected by chance (94%) and to laboratory results (96.7%). However, the detection rate for guilty examinees (42%) was much lower than expected from the results of eight laboratory studies (88.2%) cited by Lykken (1988b). One possible explanation for this difference is that in an experimental setup the experimenter has complete control over subjects' knowledge and can guarantee that all guilty subjects are aware of the relevant information. In real life, however, certain aspects of the crime may be overlooked in the excitement of the event. Moreover, suspects are rarely tested immedi-

We would like to thank Gershon Ben-Shakhar, Murray Kleiner, and three anonymous reviewers for their helpful comments.

Correspondence concerning this article should be addressed to Eitan Elaad, Division of Identification and Forensic Science, Israel National Police Headquarters, Jerusalem 91906, Israel.

ately after committing the criminal act. Typically, they are tested days, weeks, and sometimes months after the crime. Some critical items may be forgotten during this period.

Another factor suggested by Elaad (1990) to account for the large number of false-negative outcomes was the use of a single measure as an indicator of accuracy. Elaad suggested that the accuracy of the GKT in the field may increase with additional physiological measures, such as respiration line length (RLL). This index is scored by measuring the length of the curvilinear patterns recorded by the polygraph respiration pen. The shorter the line, the stronger the response. Actually, the RLL combines two respiration measures that have been tested in laboratory studies: decrease in respiration amplitude (RA) and increase in respiration cycle time (RCT).

Thackray and Orne (1968) compared eight physiological measures for their detection efficiency in the GKT. Among these, RA and RCT constituted the respiration index and changes in SRR and skin potential response (SPR) were the electrodermal measures. Thackray and Orne concluded that the two electrodermal measures were more efficient than all other measures in detecting guilty knowledge. Podlesny and Raskin (1978) compared the two respiration measures with skin conductance response (SCR) amplitude and the SPR. Their results were in complete agreement with those of Thackray and Orne. Cutrow, Parks, Lucas, and Thomas (1972) compared a total of nine physiological measures, including RA, RCT, palmar galvanic skin response (GSR_p) and volar-forearm galvanic skin response (GSR_v). The GSR_p produced the best discrimination between relevant and irrelevant items in the GKT paradigm. A combined score, based on six measures including RA, RCT, and GSR_p, produced the most efficient discrimination, which was significantly better than the best single measure.

The RLL may be sought as a combined measure of both RA and RCT. As such, the RLL provides a global score of respiratory suppression, which was indicated by Lynn (1966) as a component of the orienting response. It seems that the RLL is entitled to be examined for its own detection efficiency rather than assessed by that of each of its components. Timm (1982b) was the first to demonstrate the efficiency of the RLL in a mock-crime GKT experiment. Using Lykken's (1959) scoring procedure on 270 guilty subjects, Timm reported that the RLL was more efficient than SRR amplitude in detecting guilty knowledge. Furthermore, the highest detection rate was obtained with a combined index that was composed of SRR amplitude, SRR maximum height, and the RLL. Timm (1984) demonstrated that the RLL also resisted habituation. In the first two GKTs, the RLL did not surpass SRR amplitude in detection efficiency. Only beyond this stage, when SRRs had been diminished because of habituation, did the RLL index exhibit a better detection efficiency. Timm suggested that, although the detection efficiency of the RLL appears to reflect the signal value of the relevant item, it is also possible that the RLL monitors something other than a simple orienting response. In a more recent study, Timm (1987) reported 50% (17 of 34) accuracy in identifying guilty subjects with both the SRR and the RLL.

An independent replication of Timm's results was carried out by Elaad (in press), using an experimental GKT paradigm. In this study, 75% (24 of 32) of the guilty subjects were correctly identified with SRR amplitude, whereas only 47% (15 of 32) of the guilty subjects were detected by the RLL. However, the results obtained for the RLL index were not at all poor when compared with chance expectancy, which was less than 8%. Elaad and Timm agreed completely on the RLL detection rate. Thus, the SRR index accounts for the different results. Elaad obtained a substantially better SRR efficiency value, which corresponded to the mean accuracy rates (84%) reported for guilty subjects in 10 GKT experimental studies reviewed by Ben-Shakhar and Furedy (1990). Timm's rather poor efficiency should be considered as reflecting only the lower bounds for the efficiency of the electrodermal GKT in experimental setups.

Kircher and Raskin (1988) examined the RLL's accuracy within a CQT experiment and reported that the RLL was the second most useful measure (inferior only to SCR amplitude). Timm (1982a) examined the RLL within a real-life CQT context. On the basis of verification by confession, Timm selected 46 verified charts. Half of the sample was drawn from innocent examinees and the other half from guilty examinees. Timm reported that 17 of the 23 guilty subjects (74%) had their smallest RLL response on one of the relevant questions, whereas 16 of 23 (70%) innocent subjects had their smallest RLL response on a control question. Timm concluded that the RLL is a highly valid index for differentiating between guilty and innocent examinees within the CQT context.

However, some examinees are more responsive in one physiological channel than in the others. Therefore, higher detection efficiency levels would be expected if classification of guilty and innocent examinees were based on multiple physiological indices. Although the RLL is a promising index in the psychophysiological detection of deception, knowledge about its relative efficiency is insufficient. For example, this measure has never been compared with SRR in a real-life GKT situation. The main purpose of the present study was to assess the efficiency of the objectively measured RLL in real-life criminal examinations using the GKT. For this purpose, the present study was designed as a constructive replication of Elaad's (1990) study, which provided an opportunity to compare the relative efficiency of the electrodermal and respiration measures in detecting guilty knowledge. Furthermore, because of unshared variance, the integration of the two measures was likely to enhance detection efficiency in the GKT.

Method

Sample

A sample of 80 actual GKT criminal polygraph records, taken from police investigations conducted between 1985 and 1991, was drawn from the pool of verified GKT polygraph tests of the Israel Police Scientific Interrogation Unit.[1] The sampling was random with the exception that half of the records (with a total of 76 GKT questions) were drawn from a pool of about 90 polygraph records of verified deceptive examinees and the other half (with a total of 68 GKT questions) were drawn from a pool of about 100 innocent examinees. Each record con-

[1] Part of the present study's results were reported at the North Atlantic Treaty Organization's Advanced Study Institute on Credibility Assessment, Maratea, Italy, June, 1988. In that report, only 40 GKT records conducted between 1985 and 1988 were used.

tained from one to six GKT questions ($M = 1.80$, $SD = 0.91$). After the buffer, each question presented from four to eight multiple-choice items ($M = 6.19$, $SD = 0.66$). Each question was repeated two, three, or four times ($M = 3.25$, $SD = 0.53$). The serial position of the critical item in the set ranged from the second position to the last position (the buffer was in the first position).

GKT questions in which numbers were presented (e.g., amount of money, number of credit cards, etc.) presented a mixture of GKT and peak of tension (POT) procedures. In the POT, the set of items is arranged in a certain sequence known to the examinee. For example, in a POT test, the numbers from 1 to 6 may be presented in ascending order, whereas in the GKT, they may be randomly ordered. Furthermore, in POT tests the relevant item is inserted in an arbitrary serial position and usually will not appear as one of the first three items. Guilty examinees are expected to show a gradual increase in response amplitude until the relevant item is presented. At that point, a sharp drop in response amplitude is usually observed. The discontinuity of the response pattern is included in the criteria of guilt.

In the present study, 20 of the 76 questions presented to guilty examinees introduced such a mixture. In the first repetition, the numbers were presented in ascending order. In the second repetition, the set of items was presented in the reverse order. The third and fourth repetitions were of the GKT nature, and the examinee was unaware of the serial position of the critical item. In any case, the discontinuity of the response amplitude after the relevant item was not included as an indicator of guilt. Therefore, the term GKT is appropriate for describing the whole sample.

Verification of guilt and innocence was based on the confession of the person who had committed the crime, on the assumption that the confession confirmed the guilt of the confessor and exonerated other examinees from involvement in the investigated crime. The confession criterion has its shortcomings, of which the vulnerability to sample selection biases is the most prominent (Elaad & Schahar, 1985; Iacono, 1991; Patrick & Iacono, 1991). There may be a relationship between polygraph results and the probability that an examinee will confess, a possibility that is more relevant for results concerning guilty examinees. Furthermore, confessions are not infallible and false confessions can occur.

In spite of the possible biases, reliance on confessions is a commonly accepted method in field studies of polygraph validity (Elaad, 1990; Horvath, 1977; Kleinmuntz & Szucko, 1984). As long as no other effective actual truth criterion is available, there is no practical alternative to using this criterion to study factors affecting the application of GKT procedures in actual field conditions. Moreover, because the goal of the present study was to compare two measures with the same subject population, the sample bias effect was less relevant.

All the GKT records were of polygraph examinations conducted by 16 trained field examiners. All examiners had university degrees (from B.A. up to Ph.D) in psychology or closely related fields (criminology, social work). At the time of the test, the examiners had from 0.5 to 16 years of experience in conducting polygraph tests ($M = 7.85$, $SD = 4.88$). All tests but one were administered after a standard CQT conducted by the same examiner.

While conducting the GKT, the examiners were aware of background information about the case and of the preceding CQT result. Under these circumstances, knowledge of the critical item in each question could have introduced an experimenter expectancy bias (see Rosenthal, 1976). Such a bias could contaminate the tests' outcomes and could flaw estimates of the method's efficiency. Elaad (1990) demonstrated that the false-positive errors in field GKTs were not more frequent than expected by chance when a decision rule that is frequently used in experimental GKT studies (Lykken, 1959) was used. This may indicate that the experimenter bias was not detrimental to the outcomes of the innocent examinees. Nevertheless, the danger that expectations could flaw the results was considered in the present study.

Apparatus and Procedure

The GKTs were conducted in small quiet rooms, which contained a table with a built-in polygraph, two chairs, a carpet on the floor, and bare walls. The polygraphs used were Lafayette field models. Each polygraph recorded changes in respiration, electrodermal responses, and cardiovascular activity. Respiration was recorded by two pneumatic rubber tubes positioned around the thoracic area and abdomen. Electrodermal recording was conducted with a standard field procedure, which differs in several aspects from experimental laboratory procedures (e.g., skin resistance is measured instead of skin conductance, stainless steel electrodes are used instead of silver-silver-chloride electrodes, and electrode paste is not used). The electrodes were attached to the volar side of the index and fourth fingers of the examinee's left hand. Cardiovascular activity was recorded with a pneumatic pressure cuff positioned around the upper portion of the examinee's right arm. The cuff was inflated to a pressure between 40 and 50 mm Hg.

Data Acquisition

The largest SRR amplitude within 10 s of the presentation of each GKT alternative (which reflects the maximal decline in skin resistence) was measured in millimeters with a ruler on the pattern recorded by the polygraph SRR pen. Scoring was conducted by an examiner who did not have prior knowledge of the correct alternative, to prevent any further contamination of the results. The response to the first item in each set of items served as a buffer to discharge the examinee's tendency to react strongly to the initial item, and was excluded from measurement. In cases of external disturbances (movements, deep breath, noise, etc.), which were indicated on the polygraph chart by the original examiner, the item containing the artifact was excluded from the measurment. The whole set was excluded when the disturbance occured during presentation of the relevant item.

To clarify the data acquisition procedure, the following is an example of a GKT test that was included in the present sample. On May 15th, 1990, at night, a National Panasonic video camera was stolen from a lecture room in a military base. Two soldiers who were on duty that night were suspected of the crime. Both were brought in for a polygraph test. Several days after the test, one of the suspects confessed and returned the stolen camera.

In the GKT, after denying knowledge of the correct answer, the culprit was asked the following GKT question:

Regarding the brand of the stolen camera, do you know that
1. The brand of the camera was JVC?
2. The brand of the camera was Sony?
3. The brand of the camera was Panasonic? (*This was the critical item.*)
4. The brand of the camera was Grundig?
5. The brand of the camera was General Electric?
6. The brand of the camera was Sanyo?
7. The brand of the camera was Phillips?

In the first repetition, the critical item appeared in the third position, and the following SRR scores were computed: 47 for Item 2, 15 for Item 3, 18 for Item 4, 29 for Item 5, 2 for Item 6, and 44 for Item 7. Item 1 was the buffer. In the second repetition, the set of items was presented in the reverse order, and the following SRR scores were computed: 1 for Item 6, 88 for Item 5, 16 for Item 4, 132 for Item 3, 1 for Item 2, and 15 for Item 1. Item 7 was the buffer. In the third repetition, the items were presented in a mixed order, and the following SRR scores were com-

puted: 89 for Item 1, 129 for Item 3, none for Item 5 (because of a disturbance), 6 for Item 6, 26 for Item 2, and 1 for Item 4. Item 7 was again the buffer.

Respiration line length for each item in a set was computed with a Microvax II computer. Polygraph paper respiration recordings were transformed into a video image by means of a video camera and conveyed to a digital frame store card (Imaging Technology FG 100), with which the video image was digitized. A digital video image consists of a two-dimensional array of numbers called *picture elements* or *pixels*. The array dimensions correspond to the two-dimensional space of the paper recording, and the pixel value corresponds to the degree of luminescence at a given coordinate. Thus, the dark line of the respiration pattern on the light paper is represented by a series of array coordinates containing low pixel values. An edge-detection algorithm was applied to the digital array to identify this series of coordinates, and the distance between each consecutive coordinate pair was calculated. The average of distances between coordinate points for 15 s from stimulus onset was the RLL index.

Timm (1982a) noted that the line length of the curvilinear respiration pattern is disproportionally affected by the starting point of measurment. Thus, measuring from a point in the rapidly ascending inspiration line (line B in Figure 1) and from a point at the end of the expiration line, where changes are relatively slow (line A), would result in different line-length values for equal time intervals.

To deal with this problem, five length measures were conducted for each respiration pattern. The first started with the pixel that represented the proximal stimulus onset, the second started with the next pixel, and the fifth sampling started with the fifth pixel after stimulus onset. The mean RLL of the five samplings determined the index, thus moderating any possible extreme result due to rapid changes within the stimulus onset area. For example, when the measuring procedure was applied on the respiration line in Figure 1 (from point 1 to point 2), the following results were obtained: 2.41, 2.36, 2.30, 2.29, and 2.29, with a mean of 2.33 (each score represents the average number of length units computed for each pixel on the horizontal axis). This demonstrates that an averaging procedure provides a better estimate of the curvilinear length than does each individual measurement within the stimulus onset area.

Data Quantification

The responses to the various items in each set were rank ordered. The response that yielded the largest amplitude (when the SRR measure was used) was ranked first, the second largest response was ranked second, and so forth until all items in the set were assigned the appropriate rank. RLL scores for each item in the set were ranked such that the shortest mean length in the set was ranked first and the longest was ranked last.

Because the sizes of GKT sets vary (it turned out that, excluding the buffer, the number of alternative items per question ranged from four to eight), it is difficult to compare the rankings. For example, an item ranked second in an eight-item set should weigh more than an item ranked second in a four-item set. Therefore, some sort of correction for the ranking with relation to the set size was needed. Following Elaad (1990), an R/N ratio for each set was computed (R stands for the rank assigned to the relevant item's response, and N represents the number of items in the set, excluding all missing data).

On the basis of the R/N ratio, a detection score (Y) was defined for each question in the following way: $Y = 2$ if $R/N < .25$; $Y = 1$ if $.25 < R/N < .4$; and $Y = 0$ if $R/N > .4$. The .25 cutoff point was used to ensure that a response to the critical item ranked first would not receive a score less than 2, even in a four-item set. The .40 cutoff point was selected to resemble Lykken's (1959) second cutoff point, which consisted of all second largest responses in a five-item set. In our example, the scores computed for the SRR measure were 0, 2, and 2, for the three respective repetitions.

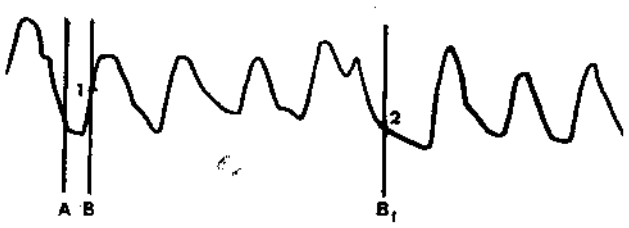

Figure 1. Curvilinear respiration tracings.

Results

Question Analysis

Each GKT series, with all the alternatives, was presented to the examinee between two and four times. The individual question scores were computed by summing the assigned category scores for each repetition. For example, a perfect deception score for a question repeated three times would be 6 (3 repetitions × 2), and a perfect truthfulness score would be 0. The questions were classified into three categories according to the following decision rules: Every question that yielded a sum of scores greater than the number of its respective repetitions was classified in the guilty knowledge indicated (GKI) category; a question that yielded a sum of scores identical to the number of its repetitions was classified in the inconclusive category; and all questions that yielded sums less than the number of their repetitions were classified as no guilty knowledge indicated (NGKI). These rules ensured that a question would be classified GKI only if a score of 2 was obtained in at least one of its repetitions. With respect to our example, the SRR question score was 4. Because the question was repeated three times, the responses were classified in the GKI category.

If GKT questions are properly constructed, the different items within each question should be entirely equivalent for an uninformed innocent examinee. Under such conditions, it can be assumed that the distribution of the ranks assigned to the different items will be rectangular, and the expected probability distribution of the scores obtained by the innocent examinee can be constructed for a given number of alternative items and repetitions. The expected probabilities for a given decision (GKI, NGKI, inconclusive), based on the number of questions, repetitions, and alternative items per questions, are presented in Table 1. There were slight differences in the expected relative frequencies computed for the SRR and RLL measures. These were due to the differential exclusion of unquantifiable responses in the two measures. The comparisons (across repetitions) of the expected probabilities and the actual decisions that were made on the basis of the SRR and RLL measures clearly indicate that the obtained results are not different from those expected.

Table 1 also presents the index of specificity, which was computed as the ratio NGKI/(GKI + NGKI) among innocent examinees. High specificity means that the test was successful in preventing false-positive errors. Across repetitions, the ob-

Table 1
Obtained Relative Frequencies and Expected Probabilities of Guilty-Knowledge-Test Decisions Made for Questions Presented to Innocent Examinees

Number of repetitions/measure	*N*	Obtained				Expected			
		GKI	Inc.	NGKI	Specificity	GKI	Inc.	NGKI	Specificity
Two	5								
SRR		0	0	1.000	1.000	.079	.259	.662	.893
RLL		0	.200	.800	1.000	.067	.231	.702	.913
Combined		0	.200	.800	1.000				
Three	49								
SRR		.041	.082	.878	.955	.083	.114	.803	.906
RLL		.041	.082	.878	.955	.082	.114	.804	.906
Combined		.082	.143	.776	.904				
Four	14								
SRR		0	.286	.714	1.000	.059	.114	.827	.933
RLL		.143	.143	.714	.833	.058	.110	.832	.935
Combined		.143	.357	.500	.778				
Across repetitions	68								
SRR		.029	.118	.853	.967	.078	.124	.798	.911
RLL		.059	.103	.838	.934	.076	.122	.802	.913
Combined		.088	.191	.721	.891				

Note. *N* = numbers of questions; GKI = guilty knowledge indicated; Inc. = inconclusive; NGKI = no guilty knowledge indicated; Specificity = NGKI/(GKI + NGKI).

tained specificity values computed for both SRR and RLL measures were not smaller than the expected chance specificity values.

Unlike the case of the innocent examinees, there is no a priori model by which the probabilities of the different decisions (GKI, inconclusive, NGKI) could be predicted for guilty examinees. Therefore, no expected probabilities are presented in Table 2. A measure of the test's sensitivity, defined as GKI/(GKI + NGKI) among guilty suspects, was computed and is displayed in Table 2. A high sensitivity level should be achieved only if the relevant alternative has a special meaning for the guilty examinee and thus elicits consistently larger responses. Lykken (1981) suggested that a properly constructed GKT should yield 80% correct detections of guilty subjects. Nevertheless, the obtained sensitivity for both measures in this study was substantially smaller (see Table 2). According to psychometric considerations the sensitivity of each GKT question should increase with repetition. However, questions that were repeated four times exhibited lower sensitivity than did questions repeated three times for both the SRR and RLL measures. This corresponds to previous results obtained in a real-life GKT study (Elaad, 1990). It seems that the polygraph examiner's de-

Table 2
Relative Frequencies of Guilty-Knowledge-Test Decisions Made for Questions Presented to Guilty Examinees

Number of repetitions/measure	*N*	GKI	Inconslusive	NGKI	Sensitivity
Two	2				
SRR		0	.500	.500	.000
RLL		.500	.500	0	1.000
Combined		.500	.500	0	1.000
Three	44				
SRR		.477	.159	.364	.568
RLL		.477	.091	.432	.525
Combined		.614	.136	.250	.711
Four	30				
SRR		.467	.100	.433	.519
RLL		.400	.067	.533	.428
Combined		.600	.133	.267	.692
Across repetitions	76				
SRR		.461	.145	.395	.538
RLL		.447	.092	.461	.493
Combined		.605	.145	.250	.708

Note. *N* = no. of questions; GKI = guilty knowledge indicated; NGKI = no guilty knowledge indicated; Sensitivity = GKI/(GKI + NGKI).

cision to repeat a given question four times may be related to the results of previous presentations of that question. Thus, the fourth repetition was more likely in those cases in which three repetitions had not produced clear results. To substantiate this explanation, the 30 questions that were repeated four times were recomputed with the fourth repetition excluded. Results indicated similar low sensitivity values (.478 and .417 for SRR and RLL indexes, respectively).

If unshared variance exists, the combination of both the RLL and SRR measures may improve detection efficiency. To examine this hypothesis, we proposed a new decision rule that integrated both measures into a single index. According to the new rule, an indication of guilty knowledge by either SRR amplitude or RLL should lead to a GKI decision. An inconclusive indication by one or both measures should lead to a final inconclusive decision. An NGKI decision should be made when both measures indicate no guilty knowledge. The new decision rule trades false-negative decisions with true negatives. In other words, the rule is designed to decrease false-negative errors at the risk of obtaining a slight reduction in the number of correct decisions about innocent examinees.

The combined index is presented in Tables 1 and 2. Across repetitions, this combination enhanced the correct GKI decision rate to 60.5% and sensitivity to 70.8% (Table 2). The relative frequency of false-positive errors was set at 8.8%, and specificity decreased to 89.1% (Table 1). For innocent examinees, the expected probabilities of the different decisions made according to the combined measure can be estimated only on the basis of an assumption of independence. This assumption cannot be applied for the two physiological measures. Therefore, no expected probabilities for the combined index are presented in Table 1.

Examinee's Analysis

To compute a global score (S) for each examinee, the question scores were summed up again. Questions that were classified GKI, inconclusive, or NGKI were assigned scores of 2, 1, and 0, respectively. Thus, a perfect deception score for an examinee presented with only one GKT question would be 2 (1 question × 2), and a perfect truthfulness score would be 0. The following decision rule was used to classify examinees: A GKI classification was made if $S > Q$, where Q stands for the number of questions presented to the examinee. According to the rule, a score of at least 4 was needed to classify as GKI an examinee presented with three questions. An inconclusive decision was reached if $S = Q$, and an NGKI decision was made whenever the score computed for the examinee was less than the number of questions presented. This ensured that an NGKI classification could not be made unless at least one question was assigned a score of 0 and that a GKI classification was possible only if at least one question was assigned a score of 2. The relative frequencies of GKT decisions made for innocent and guilty examinees are presented in Tables 3 and 4.

Assuming that for an innocent examinee the ranks assigned to the different items have a rectangular distribution, it is possible to extract the expected probability that an innocent examinee will be classified as guilty by the present decision rule. The expected probabilities of classifying an innocent examinee as guilty, innocent, or inconclusive are described in Table 3 as a function of the number of GKT questions, repetitions, and items presented to the examinees.

Table 3 reveals further that, of 40 innocent examinees, only 1 (2.5%) was incorrectly identified as possessing guilty knowledge for each of the two measures. This false-positive error rate corresponds to the expected error rate when a rectangular distribution of the ranks is assumed.

With regard to guilty examinees (Table 4), 16 of 40 guilty examinees (40%) were correctly assigned to the GKI category when the SRR measure was used, and 17 of the 40 guilty examinees (42.5%) were correctly assigned to the GKI category when the RLL measure was used. The false-negative error rates computed for the SRR (35%) and RLL (37.5%) measures also did not differ.

The combined index (both SRR and RLL measures) was computed and yielded enhanced detection of guilty examinees (62.5%; 25 of 40) while keeping the false-positive error frequency relatively low (5%; 2 of 40). As indicated by McNemar's chi-square test for correlated proportions, the combined index detection rate is significantly better (at the .05 level) than the SRR detection rate, $\chi^2(1, N = 40) = 7.1$, and the RLL detection rate, $\chi^2(1, N = 40) = 6.1$.

Regarding innocent examinees, the obtained specificity values computed across questions for the SRR and RLL measures (97.4% and 97.2%, respectively) do not differ from the expected specificity values (94.5% and 94.6%, respectively). Furthermore, they resemble the specificity value (97.9%) obtained for the SRR measure in a previous field study (Elaad, 1990) and the overall specificity (96.7%) found in eight laboratory studies of the GKT (Lykken, 1988b).

The overall sensitivity in the eight laboratory studies cited by Lykken (1988b) indicated that the GKT was very sensitive (88.2%). However, a relatively low sensitivity value (50%) was obtained in Elaad's (1990) previous field study. The present sensitivity values (53.3% and 53.1% for the electrodermal and respiration measures, respectively) are the same. The sensitivity value computed for the combined index (75.8%) is larger because of the unshared variance and the decision rule that traded higher sensitivity for somewhat lower specificity.

The test's sensitivity was expected to increase with additional test questions. Table 4 exhibits such a gradual increase in sensitivity for the combined index. However, as tested with the normal approximation to the binomial distribution, the sensitivity computed for the two-question tests did not differ significantly from the sensitivity computed for the one-question tests at the .05 level ($Z = 0.65$). Similar results were obtained when the sensitivity computed for the tests involving three or more questions was compared with the sensitivity of the two-question tests ($Z = 0.60$). In comparison, Elaad (1990) obtained the largest sensitivity for the two-question tests, with a gradual decrease in sensitivity with the addition of further questions.

The use of a predefined decision rule (Lykken, 1959) may have limited the overall accuracy rate. Therefore, a second method was used to analyze the results to ensure that conclusions would not depend entirely on an arbitrary decision rule. As in previous studies (e.g., Ben-Shakhar, 1977; Elaad & Ben-Shakhar, 1989; Lieblich, Ben-Shakhar, & Kugelmass, 1976), signal detection measures were used to analyze the data of the present study.

Table 3
Obtained Relative Frequencies and Expected Probabilities of Guilty-Knowledge-Test Decisions Made, Across Questions, for Innocent Examinees

Number of questions/measure	*N*	Obtained				Expected			
		GKI	Inc.	NGKI	Specificity	GKI	Inc.	NGKI	Specificity
One	19								
SRR		0	.053	.947	1.000	.077	.124	.799	.912
RLL		.053	.210	.737	.933	.077	.124	.799	.912
Combined		.053	.263	.684	.928				
Two	14								
SRR		.071	.071	.857	.923	.023	.131	.846	.974
RLL		0	0	1.000	1.000	.022	.128	.850	.975
Combined		.071	.071	.857	.923				
Three	7								
SRR		0	0	1.000	1.000	.025	.052	.923	.974
RLL		0	0	1.000	1.000	.023	.049	.928	.976
Combined		0	0	1.000	1.000				
Across questions	40								
SRR		.025	.050	.925	.974	.049	.114	.837	.945
RLL		.025	.100	.875	.972	.048	.112	.840	.946
Combined		.050	.150	.800	.941				

Note. *N* = number of examinees; GKI = guilty knowledge indicated; Inc. = inconclusive; NGKI = no guilty knowledge indicated; Specificity = NGKI/(GKI + NGKI).

First, the responses were standardized to compare those obtained from different examinees in different questions and repetitions. Each response was transformed into a standard score relative to the mean and standard deviation of the examinee's responses within each repetition. Second, the *Z* scores of the responses to relevant items were averaged within each examinee, across questions and repetitions, yielding new indexes of detection for SRR and the RLL. To observe the shared and unshared variance between SRR and the RLL, we computed a Pearson correlation between the two *Z* scores. The obtained coefficient of .373, though significant at the .01 level, yielded a common variance of .139. This indicates that the unshared variance is considerable.

Next, the mean of the two *Z* scores was computed for each examinee to serve as a new combined index. Finally, the frequency distributions of the mean *Z* scores were computed for guilty and innocent examinees for each physiological measure and for the combined index. These distributions are displayed in Table 5.

Table 5 is useful for comparing the distribution of the mean

Table 4
Relative Frequencies of Guilty-Knowledge-Test Decisions Made, Across Questions, for Guilty Examinees

Number of questions/measure	*N*	GKI	Inconclusive	NGKI	Sensitivity
One	17				
SRR		.353	.176	.470	.429
RLL		.412	.118	.470	.467
Combined		.647	.059	.294	.688
Two	13				
SRR		.385	.385	.230	.625
RLL		.308	.308	.385	.444
Combined		.615	.230	.154	.800
Three or more	10				
SRR		.500	.200	.300	.625
RLL		.600	.200	.200	.750
Combined		.600	.300	.100	.857
Across questions	40				
SRR		.400	.250	.350	.533
RLL		.425	.200	.375	.531
Combined		.625	.175	.200	.758

Note. *N* = number of examinees; GKI = guilty knowledge indicated; NGKI = no guilty knowledge indicated; Sensitivity = GKI/(GKI + NGKI).

Table 5
Frequencies and Cumulative Relative Frequencies of Guilty and Innocent Examinees' Mean Z Scores on the Relevant Itrems, Computed with the SRR, RLL, and Combined Measures

	Guilty			Innocent			Cumulative % of guilty decision					
							SRR		RLL		Combined	
Z scale	SRR	RLL	Combined	SRR	RLL	Combined	G	I	G	I	G	I
1.40+	4	3	0	0	0	0	10	0	8	0	0	0
1.20	3	2	3	0	0	0	18	0	13	0	8	0
1.00	2	3	3	1	1	0	23	3	20	3	15	0
0.90	0	3	5	0	0	0	23	3	28	3	28	0
0.80	5	3	1	0	0	0	35	3	35	3	30	0
0.70	3	3	4	0	1	0	43	3	43	5	40	0
0.60	5	1	5	0	2	0	55	3	45	10	53	0
0.50	2	5	3	3	1	1	60	10	58	13	60	3
0.40	0	2	3	0	2	2	60	10	63	18	68	8
0.30	4	4	2	1	1	0	70	13	73	20	73	8
0.20	2	1	2	4	3	2	75	23	75	28	78	13
0.10	3	2	1	4	2	5	83	33	80	33	80	25
0.00	2	2	2	1	4	7	88	35	85	43	85	43
−0.10	0	0	2	3	4	5	88	43	85	53	90	55
−0.20	1	2	4	7	3	5	90	60	90	60	100	68
−0.30	1	1	0	5	3	2	93	73	93	68	100	73
−0.40	0	1	0	4	2	8	93	83	95	73	100	93
−0.50	1	1	0	2	5	0	95	88	98	85	100	93
−0.60 or less	2	1	0	5	6	3	100	100	100	100	100	100

Note. G = guilty; I = innocent.

Z scores computed for guilty examinees with the distribution of the mean Z scores computed for innocent examinees. Each point (Z_i) along the mean response scale determines a hit-rate value and a false-positive value. The hit rate was defined as the proportion of cases from the guilty-examinee distribution that elicited a mean Z score higher than Z_i. The false-positive rate was defined as the proportion of cases from the innocent-examinees distribution that elicited a Z score higher than Z_i. Table 5 reveals that if a false-positive error rate of up to 5% can be tolerated, the Z_i cutoff for the SRR can be set at 0.60. Assigning subjects with a mean SRR Z score greater than or equal to 0.60 to the GKI group yielded a hit rate of 55% (22 of 40). When the RLL measure was used, the Z_i cutoff was set at 0.70 and the hit rate was 43% (17 of 40). The cutoff for the combined index was set at 0.50, and 60% (24 of 40) of the guilty examinees were correctly classified.

Finally, a receiver operating characteristic (ROC) curve was generated for each individual physiological measure and for the combined index. The ROC curves were generated by comparing the distribution of the mean Z scores computed for guilty examinees with the distribution of the mean Z scores computed for innocent examinees across all Z_i values. The area under each of the three ROC curves was computed (see Table 6), resulting in a statistic that assumes values between 0 and 1. An area of 0.5 indicates that both distributions are undifferentiated. An area of 1 indicates that there is no overlap between the two distributions.

Bamber (1975) showed that the area under an ROC curve has an asymptotic normal distribution. He described a method for estimating the variance of the area statistic and for computing confidence intervals for the true area. Using Bamber's method, a 95% confidence interval was computed for each area. The areas under the ROC curves, as well as the corresponding 95% confidence intervals, are presented in Table 6.

Table 6 reveals that both the RLL and SRR amplitude were good indicators of guilty knowledge (the area in both are significantly better than a chance area of .50). The combined index seemed to slightly improve detection efficiency.

Discussion

Four clear effects emerge from the results of the present study. First, the RLL index was an efficient index for the detection of guilt. This result extends the previous experimental findings of Timm (1987) and Elaad (in press) to the actual criminal GKT context. The practical implications are that the results of the present study may pave the way for the application of the objectively measured and well-integrated RLL in real-life criminal cases.

Second, SRR amplitude and the RLL were equally accurate

Table 6
Area Under the ROC Curve and a 95% Confidence Interval Computed for SRR, RLL, and Combined Mean Z Scores for the Relevant Items

Measure	Area	95% confidence interval
SRR amplitude	.833	.740–.926
RLL	.806	.711–.902
Combined index	.879	.806–.953

Note. ROC = receiver operating characteristic.

in discriminating between guilty and innocent examinees in the context of actual GKTs. This seems to be in contrast with the results of experimental GKT studies, which have consistently indicated that electrodermal measures are more efficient than respiration measures in the detection of guilty knowledge. Thackray and Orne (1968), for example, obtained a clear advantage for electrodermal measures over RA and RCT. Podlesny and Raskin (1978) conducted a mock-crime experiment and obtained better detection with SCRs than with RA and RCT. Elaad (1987) reported that the objectively measured SRR amplitude proved to be more efficient than the polygraph examiner's evaluation of respiration responses. However, neither the evidence about RA and RCT efficiency, nor the evidence about subjectively evaluated respiration responses, can be simply applied to the integrated and objectively measured RLL. The only evidence about the superiority of the SRR over the RLL in the experimental context emerged from a study conducted by Elaad (in press). In contrast, two other laboratory GKT studies in which the RLL was compared with SRR amplitude (Timm, 1982b, 1987) indicated that the RLL index is as efficient as the SRR measure. A closer inspection of the results reveals that both Elaad and Timm agreed on the RLL detection rate. Elaad, however, reported a substantially higher SRR efficiency (75%) than did Timm (about 50%). It has been suggested that Timm's rather poor SRR efficiency should be regarded as reflecting only the lower bounds for the measure's efficiency in the experimental setup.

The conflict between the SRR's superiority over the RLL in Elaad's (in press) laboratory study and the equal detection efficiency of both in the present field study can be resolved by the suggestion that in field conditions an increase in RLL efficiency and a drop in SRR efficiency, relative to the experimental results, is observed.

There is evidence that some guilty examinees do not notice or remember all the facts of the crime available to the investigator. Thus, for example, a guilty examinee who was tested regarding the color of the scarf with which he strangled the victim didn't respond differentially to the relevant and neutral items. It is possible that for this examinee the color of the scarf was an insignificant detail and thus was overlooked or forgotten. However, the same examinee responded differentially to a similar question about the color of the blanket with which the body was covered. This pair of questions characterizes a situation sometimes encountered in the world outside the laboratory, in which the perception and retention of information depends on individual factors such as mood, personal interest in the information, retroactive or proactive misdeeds of the culprit, or the time elapsed from the crime to the test.

If, for example, the guilty suspects were aware of only 70% of the guilty knowledge items, the obtained 53% RLL and SRR efficiency corresponds to 76% efficiency under the assumption that all examinees were aware of all the relevant information. In any case, the observed 53% efficiency underestimates the actual efficiency of both the RLL and SRR.

SRR's field efficiency is less than that reported in laboratory studies. One explanation for the difference may be attributed to the subject's level of arousal. The SRR may be more sensitive than the RLL to low levels of arousal, such as that found in laboratory experiments, whereas both measures may be equally sensitive to higher levels of arousal, such as that found in real-life tests. An alternative explanation may be that implementing the GKT after a standard CQT may interfere with SRR's efficiency as an indicator of guilty knowledge. Respiration responses, which are based on voluntary attention, may be less susceptible to habituation, and therefore RLL efficiency is not affected.

The third effect that emerged from the present study's results corresponds to the integration of the two measures into a new one. The combined measure tended to increase detection efficiency. A large number of false-negative errors was obtained for the RLL measure as well as for the SRR, indicating that the false-negative phenomenon is not confined to a single physiological index. To decrease the false-negative error rate, the combined index, which trades false-negative decisions with true negatives, was applied. When this combined index was used, the correct decision rate for guilty examinees increased from either 40% (SRR) or 42.5% (RLL) to 62.5% (see Table 4). This trade resulted in the increase of false-positive decisions from 2.5% to 5% (see Table 3). The confirmed suggestion that the combined index would further improve detection is consistent with previous sucessful attempts to combine physiological measures (Cutrow et al., 1972). Moreover, combining physiological measures is routine in field practice. Effort should be invested in searching for additional physiological measures that may improve detection efficiency even further.

Fourth, the SRR identification rates for both guilty and innocent examinees are in complete agreement with those obtained by Elaad (1990) in a previous field study. This lends support to the external validity of these results. The identification rates obtained in both field studies agree with those of laboratory studies with respect to innocent examinees.

Regarding guilty examinees, the detection rates reported in both field studies (Elaad, 1990, and the present study) are considerably lower than those reported in various experimental studies (Davidson, 1968; Elaad & Ben-Shakhar, 1990; Elaad, Bonwitt, Eisenberg, & Meytes, 1982; Giesen & Rollison, 1980; Lykken, 1959, 1960; Podlesny & Raskin, 1978). Previous findings in experimental situations (Elaad & Ben-Shakhar, 1989; Gustafson & Orne, 1963) have demonstrated that motivating subjects to avoid detection leads to enhanced detection of guilty knowledge. This implies that high levels of motivation, such as typically characterize actual polygraph examinations in real-life settings, should contribute to better detection rates than the level of detection obtained in experimental situations. From this perspective, the present results are disappointing.

As indicated earlier, the large false-negative error rate obtained in the present study and in the previous field study (Elaad, 1990) seems to reflect the main shortcoming of the GKT—Some guilty examinees are unaware of some facts of the crime. It is possible that a person who commits a crime may overlook some relevant information, especially during strong excitement, or may forget it later on.

Appropriate use of the GKT procedure requires that the guilty examinee be aware of the guilty knowledge. However, it is difficult to foresee whether a guilty examinee will be aware of a certain item, even when the item seems to be salient to an objective observer. We can only stress Elaad's (1990) conclusion that, to improve detection efficiency, the most salient details possible should be selected to formulate the guilty knowledge questions while taking into consideration the following factors:

(a) How much time passed from the crime to the test? (b) Is the examinee expected to be interested in the question's content? (c) Was the relevant information actively acquired by the guilty person? (d) Are the relevant items to be prepared in advance by the culprit? (e) Is it likely that retroactive or proactive misdeeds of the culprit may interfere with the salience of the crime-relevant information? (f) Are the items presented in the test distinctive enough to be recognized by the guilty examinee? Using such a cautious approach may decrease the probabilty that the relevant item may have been overlooked by the culprit.

In the present study, the mean number of GKT questions that were used was only 1.80, whereas in the eight laboratory studies cited by Lykken (1988b), the number of GKT questions varied from 5 to 10. Table 4 presents a trend of gradual increase in sensitivity with the addition of questions, supporting the notion that the small number of GKT questions that were used may have contributed to the large false-negative error rate.

The predefined decision rule produced a low rate of false-positive errors. This indicates that the innocent examinees were not aware of the relevant information. If there were a few cases in which the information might have been leaked to the innocent examinee, the awareness of the guilty knowledge had only a marginal effect on the outcomes of the GKT. This concurs with the findings of Bradley and Warfield (1984), who addressed the problem of innocent subjects' knowledge. They concluded that the mere possession of guilty knowledge (in an innocent context) has a much weaker impact on differential responsivity than does similar knowledge in a guilty context.

In the present study, GKT records were drawn ex post facto, and the test situation could not have been planned beforehand. Thus, factors such as the examiner's awareness of the critical information while conducting the test are inherent in the situation. The low frequency of false-positive errors indicates that the examiner's knowledge of the critical information did not flaw the test's results. Similar results were obtained in Elaad's (1990) previous field study and in an earlier study conducted at the Israeli police laboratory (Elaad & Shifron, 1984), which was designed to investigate the effect of polygraphers' expectations on the responses obtained from subjects in the GKT. The conclusion was that the expectancy effect is rather marginal in the GKT paradigm.

Finally, the use of confessions as the criterion of validity can be criticized for not controlling a possible sampling bias. The probability that a subject will confess may depend on the polygraph results in that a deceptive outcome may encourage interrogation efforts to induce a confession. On the other hand, a truthful outcome may convince the police interrogator to dismiss suspicion against the subject. This may lead to an underestimation of the false-negative rate. Furthermore, one cannot exclude the possibility of false confessions. Such confessions may lead to an overestimation of the false-negative rate. Only the use of a solid criterion for truth, under highly realistic conditions, such as that used by Ginton, Daie, Elaad, and Ben-Shakhar (1982), can address this problem. However, Ginton et al.'s (1982) procedure caused a considerable dropout of guilty subjects before the test was taken.

To examine the impact of the GKT results on the confessions of guilty examinees, we divided the guilty examinees into two groups according to their time of confession. All subjects who confessed immediately after the polygraph examination or within the same day were assigned to the immediate group (N = 15). All other guilty examinees confessed afterward, so that additional interrogational factors may have been involved, thus reducing the direct link between the polygraph results and the confession. These examinees were assigned to the delayed group (N = 25). A comparison of these groups, using the combined index, revealed that the detection rate for the immediate group was 60% (9 of 15). The detection rate for the delayed group was 64% (16 of 25). However, the immediate group produced more inconclusive outcomes (33% compared with 8% for the delayed group), whereas the delayed confessors exhibited more NGKI results (28% compared with 7% for the immediate confessors). These results suggest that the danger of sampling bias of guilty examinees, in the present confession GKT study, is not as prominent as recently suggested by Patrick and Iacono (1991) with respect to the CQT.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology, 12,* 387–415.

Ben-Shakhar, G. (1977). A further study of the dichotomization theory in detection of information. *Psychophysiology, 14,* 408–413.

Ben-Shakhar, G., & Furedy J. J. (1990). *Theories and applications in the detection of deception: A psychophysiological and international perspective.* New York: Springer-Verlag.

Bradley, M. T., & Warfield, J. F. (1984). Innocence, information, and the guilty knowledge test in the detection of deception. *Psychophysiology, 21,* 683–689.

Cutrow, R. J., Parks, A., Lucas, N., & Thomas, K. (1972). The objective use of multiple physiological indices in the detection of deception. *Psychophysiology, 9,* 578–588.

Davidson, P. O. (1968). Validity of the guilty knowledge technique: The effect of motivation. *Journal of Applied Psychology, 52,* 62–65.

Elaad, E. (1987). *Psychophysiological detection in the guilty knowledge test.* Unpublished doctoral thesis, Hebrew University of Jerusalem, Israel.

Elaad, E. (1990). Detection of guilty knowledge in real-life criminal investigations. *Journal of Applied Psychology, 75,* 521–529.

Elaad, E. (in press). *Human evaluation and objective measurement in psychophysiological detection of deception.* Forensic Reports.

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

Elaad, E., & Ben-Shakhar, G. (1990). Effects of mental countermeasures on psychophysiological detection in the guilty knowledge test. *International Journal of Psychophysiology, 11,* 99–108.

Elaad, E., Bonwitt, G., Eisenberg, O., & Meytes, I. (1982). Effects of beta blocking drugs on the polygraph detection rate: A pilot study. *Polygraph, 11,* 225–233.

Elaad, E., & Schahar, E. (1985). Polygraph field validity. *Polygraph, 14,* 217–223.

Elaad, E., & Shifron, E. (1984). *Hashpaat zipiot habodek al haivchun hapsychophysiology bemivchan hapecham* [Effects of examiner expectations on psychophysiological detection in the guilty knowledge test (internal report)]. Jerusalem: Israeli Police, Scientific Interrogation Unit.

Giesen, M., & Rollison, M. A. (1980). Guilty knowledge versus innocent associations: Effects of trait anxiety and stimulus context on skin conductance. *Journal of Research in Personality, 14,* 1–11.

Ginton, A., Daie, N., Elaad, E., & Ben-Shakhar, G. (1982). A method

for evaluating the use of the polygraph in a real-life situation. *Journal of Applied Psychology, 67,* 131–137.

Gustafson, L. A., & Orne, M. T. (1963). Effects of hightened motivation on the detection of deception. *Journal of Applied Psychology, 47,* 408–411.

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

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? In J. R. Jennings, P. K. Ackles, & M. G. H. Coles (Eds.), *Advances in psychophysiolgy* (Vol. 4, pp. 201–207). London: Kingsley.

Kircher, J. C., & Raskin, D. C. (1988). Human versus computerized evaluation of polygraph data in a laboratory setting. *Journal of Applied Psychology, 73,* 291–302.

Kleinmuntz, B., & Szucko, J. J. (1984). A field study of the fallibility of polygraphic lie detection. *Nature, 303,* 449–450.

Lieblich, I., Ben-Shakhar, G., & Kugelmass, S. (1976). Validity of the guilty knowledge technique in a prisoner's sample. *Journal of Applied Psychology, 61,* 89–93.

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

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

Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist, 29,* 725–739.

Lykken, D. T. (1981). *A tremor in the blood.* New York: McGraw-Hill.

Lykken, D. T. (1988a). The case against polygraph testing. In A. Gale (Ed.), *The polygraph test: Lies, truth, and science* (pp. 111–125). London: Sage.

Lykken, D. T. (1988b). Detection of guilty knowledge: A comment on Forman and McCauley. *Journal of Applied Psychology, 73,* 303–304.

Lynn, R. (1966). *Attention, arousal and the orientation reaction.* Oxford, England: Pergamon Press.

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

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

Raskin, D. C. (1988). Does science support polygraph testing? In A. Gale (Ed.), *The polygraph test: Lies, truth, and science* (pp. 96–110). London: Sage.

Reid, J. E., & Inbau, F. E. (1977). *Truth and deception: The polygraph ("lie detector") technique* (2nd ed). Baltimore: Williams & Wilkins.

Rosenthal, R. (1976). *Experimenter effects in behavioral research.* New York: Irvington.

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

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

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

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

Timm, H. W. (1984). Significant findings attributable to electrodermal habituation effects: Artifact or essence in detection of deception research? *Journal of Police Science and Administration, 12,* 267–276.

Timm, H. W. (1987). Effect of biofeedback on the detection of deception. *Journal of Forensic Sciences, 32,* 736–746.

Received December 17, 1990
Revision received March 25, 1992
Accepted March 30, 1992 ■