Polygraph

VOLUME 27 1998 NUMBER 3

Contents

Published Quarterly

P.O. Box 8037, Chattanooga, Tennessee 37414-0037

Conditioning of Expectations in a Concealed Knowledge Test

Vance MacLaren and Michael Bradley[1]

Abstract

This study examined the possibility that detection with the Concealed Knowledge Test (CKT) is mediated by an attentional process. Students' expectations about the presentation of critical and control word items were manipulated using a series of 'training sets'. In a subsequent CKT set, presentation of items either violated or conformed to expectations acquired in the training sets. The major result was that information recognition did not appear to be the factor most responsible for selective physiological response. When subjects expected a critical word item but heard a control word, large electrodermal responses were elicited by the control. Also, expected critical words failed to elicit selective electrodermal responses at a frequency in excess of chance. Both of these results refute the notion that information recognition is the sole mechanism responsible for differential physiological response in the CKT.

Keywords: attention, concealed information test, guilty knowledge test, signal detection theory

Forensic psychophysiology exists to facilitate the prosecution of criminals and the exoneration of innocent suspects. Yet, despite numerous reports on the validity of polygraph test procedures, the practice of polygraphy continues to draw controversy in the media and in academic circles. To justify the practices of field examiners, what is needed is something more than demonstrations of the accuracy of the various techniques that are used. The legitimacy of any forensic procedure hinges upon its scientific principles and forensic psychophysiologists must demonstrate that their techniques are reasonable, not only regarding their accuracy as tests of veracity, but also with respect to theory. Theoretical advances have lagged far behind advances in recording technology and procedure, and it is this gap that is so often seized upon by lawyers and opponents of polygraphic practice. Empirically supported theory is prerequisite to gaining acceptance by scientists, laypeople, and in courts of law.

In this study, we examined the psychological basis of the Concealed Knowledge Test (CKT). Although the CKT is not widely applied in field practice, it provides a test format that is amenable to experimental testing of hypotheses which might apply equally well to more commonly used procedures, like the Comparison Question Techniques (CQT) and the Peak of Tension Test (POT).

Ever since David Lykken (1959, 1960) introduced the CKT to the psychological literature nearly four decades ago, most scientists have considered the most important factor determining whether a suspect will be found guilty or innocent by a CKT to be whether or not they possess concealed incriminating information about the crime in question. To illustrate, in a CKT such as, "The thief hid what he had stolen. Where did he hide it? Was it in the men's room... on the coat rack... in the office... on the window sill... in the locker?", only a guilty suspect would be

[1]University of New Brunswick, Canada

We wish to thank Gershon Ben-Shakhar, John Furedy, Murray Goddard, Peter McGahan, and Harold Taukulis for their advice and assistance with this project. Special thanks also to Ken Lerette for electronics and software support. Please forward correspondence concerning this paper to Vance MacLaren, 100 Leinster St., Saint John, New Brunswick, Canada, E2L-IN9, or e-mail via internet to vancem@nbnet.nb.ca.

expected to respond selectively to the correct item. An uninformed innocent suspect would be expected to respond to each of the items with equal probability. If such is the case (see Ben-Shakhar & Furedy, 1990 for a review of CKT validity), one might conclude that recognition of an item as correct is necessary for selective differential responses to be elicited. Indeed, one might go on to assume that such recognition is a discrete psychological process that is linked to autonomic activation. On the other hand, results from experimental studies have led some authors to suggest that an attentional process is the mechanism by which differential responses are generated (e.g. Elaad & Ben-Shakhar, 1989; Waid, Orne, Cook, & Orne, 1978). According to this attention hypothesis, the primary mechanism behind detection is attention, and attention may be triggered by any number of psychological events, including but not necessarily limited to recognition. It might also be triggered or augmented by deception, motivation to avoid detection, or guilt. The objective of the present study was to experimentally separate the processes of recognition and attention and to determine which, if any, is required for selective physiological response.

The Attention Hypothesis

In an early review of the scientific literature on the polygraph, Davis (1961) proposed three possible mechanisms by which the polygraph might detect guilt. These three theories were called the conflict theory, the punishment theory, and the conditioned response theory. Specific hypotheses about CKT information detection have been derived from these three models and tested experimentally, each having received some support.

According to the conflict theory, autonomic arousal is generated when a suspect tells a lie. This idea is common to all detection of deception tests, but not to Lykken's (1959) original formulation of the CKT. According to Lykken, the CKT is strictly an information detector, and information may be detected whether or not a suspect is deceptive during questioning. This has been challenged experimentally. Bradley, MacLaren & Carle (1996), Elaad & Ben-Shakhar (1989), Elaad (1993), Furedy & Ben-Shakhar (1991), Gustafson & Orne (1965), and Horneman & O'Gorman (1985) each found significant effects of verbal response on CKT detection. In each of these studies, "No' response modes (i.e. a suspect says the word "No" after each word item and is therefore deceptive to critical word items), produced rates of detection higher than when subjects were non-deceptive (i.e. subjects who remain silent, repeat the word items, or say "yes" or "maybe"). However, some researchers (Kugelmass, Lieblich & Bergman, 1967) have found less dramatic effects of deception and even in the studies that found an effect, it is difficult to attribute detection rates solely to the deception effect. In all of these studies, subjects who were informed, but not deceptive, had their information detected at rates greater than what would be expected by chance. It would seem as though a deceptive verbal response can augment physiological responses to words recognized as critical, but only information recognition is both necessary and sufficient for detection to occur.

The punishment theory proposes that the likelihood of detection increases as the suspects motivation to avoid detection increases. While some authors have found support for this notion (Gustafson & Orne, 1963; Elaad & Ben-Shakhar, 1989; Elaad & Ben-Shakhar, 1997), others have not (Bradley & Janisse, 1981; Davidson, 1968; Kugelmass & Lieblich, 1966; Lieblich, Naftali, Shmueli, & Kugelmass, 1974). It is difficult to draw conclusions from these laboratory studies, since the amount of fear or ego involvement experienced by the participants in these studies is questionable. It is certainly difficult to imagine a student feeling as concerned with a polygraph test about a mock crime or a card test as a criminal suspect facing the possibility of incarceration. Still, significant motivation effects were found in some of the studies.

According to the conditioned emotional response theory, physiological responses observed at the time of a polygraph test may be linked to fear, excitement or distress experienced at the time of the incident under investigation. The guilt

variable has been manipulated by comparing detection rates of subjects who are questioned about a mock crime with subjects who possess the same information, but who acquired it in some other way (Giesen & Rollison, 1980; Stern, Breen, Watanabe, & Perry, 1981). In both of these studies, mock crime subjects were more detectable than the informed innocent subjects. Bradley & Warfield (1984) expanded upon this design by having five groups of subjects, four of which were given crime-relevant information, but only one of which was guilty of actually acting out the mock crime. The other informed subjects either witnessed the crime, read a description of the crime, or read a non-crime related story which contained the same keywords as the mock crime story. The subjects who enacted the mock crime were detected more often than subjects in any other group, suggesting that guilt may play an important role in CKT detection. Bradley & Rettinger (1992) and Bradley, MacLaren & Carle (1996) found similar results. Other authors (Elaad & Ben-Shakhar, 1989; Iacono, Cerri, Patrick, & Fleming, 1992) have pointed out that even the relatively low detection rates of informed innocent subjects in the Bradley & Warfield study were significantly greater than chance. Results from these studies suggest that guilt may augment CKT detection of informed subjects.

When results from CKT studies on deception, motivation to avoid detection, and guilt are taken together, a pattern seems to emerge. An innocent subject who is neither deceptive, motivated, nor guilty, but who does possess concealed information may have that information detected with a likelihood somewhat greater than chance. However, the probability that they would be detected is much lower than a guilty subject who is deceptive, guilty of the crime, and motivated to avoid being caught. This, of course, is the situation that would be seen in field application of the CKT. One might expect that validity estimates obtained under laboratory conditions would be quite a bit poorer than the accuracy of the CKT in real investigations. Two field studies of the CKT under real conditions have been conducted in Israel (Elaad, 1990; Elaad, Ginton, & Jungman, 1992), and the results were surprising. Whereas Ben-Shakhar & Furedy (1990) found the average detection rate of guilty subjects across 10 laboratory studies to be 83.9% (N=248), only 41% of the 88 confirmed guilty suspects in the two field studies were found guilty by the CKT. These suspects were guilty, deceptive, and clearly motivated to avoid incrimination. What could explain these low detection rates in what would seem like optimal conditions for the CKT? Ben-Shakhar & Dolev (1996) suggested that countermeasures might have been more prevalent in the field studies than is typical of the laboratory. The susceptibility of the CKT to countermeasures has been well documented (Lykken, 1960; Elaad & Ben-Shakhar, 1991; Honts, Devitt, Winbush, & Kircher, 1996; Ben-Shakhar & Dolev, 1996). The presence of countermeasures may serve to reduce selective arousal induced by the critical items, or increase arousal to the controls.

Present research findings suggest that information recognition is the only factor that is required for CKT detection and that the other, more emotional factors merely add to the selective arousal evoked by the recognition of critical items. An alternative view is that recognition, deception, motivation, guilt, and countermeasures all contribute in separate and important ways to selective attention to the word items. From this perspective, the apparently special role of information recognition may be merely an artifact and the real main ingredient to selective physiological arousal might be selective attention. To prove this, one must separate attention from recognition and observe their effects in isolation. This was the objective of the present experiment.

Method

Participants

Participants were 72 female university students, aged 17 to 36 (median = 18) and 63 male students, aged 17 to 54 (median = 18). The participants were promised the addition of one bonus percentage point to their Introductory Psychology grade and a possible cash award of five dollars, with payment contingent upon their performance on the

polygraph test. A detailed consent form was read and signed by each participant.

Apparatus

Skin resistance was recorded using a Grass model 7 polygraph, equipped with a model 7DAG DC amplifier. UFI .83cm Ag-AgCl electrodes were attached with Velcro strips to the palmar surface of the distal phalanges of the first and second fingers on the right hand. Electrodes were filled with UFI Biogel, a .05 M NaCl conductance paste. Prior to electrode placement, the fingers were prepared using UFI Biobrade skin abrasion pads. Balance voltage and output sensitivity were individually adjusted during a three minute baseline adjustment period, such that an upward pen deflection of at least 1 cm was evoked by the subject being asked to state his or her name. Chart drive speed was 2.5 mm/s. Peripheral blood volume and respiratory measures were also taken, but are not reported here.

Pre-recorded test questions were presented using an audio cassette deck and a pair of headphones. Vocalizations made by the subject were transduced using a small microphone attached to the shirt. To mark question onset, output from the microphone and from the cassette player were fed into two voice activated switches and recorded on the marker channel of the polygraph. When triggered, the switches for tape and voice produced 2 mm pen deflections downward and upward, respectively.

The participants were seated facing a blank wall, in a polygraph chair having large padded arm rests and a beige office divider was placed between the subject and the recording equipment and operator. Air temperature inside the recording room was maintained at approximately 22°C. Noise from the recording equipment and from a personal computer inside the chamber produced an ambient noise level of approximately 60 dB.

Experimental Design

The experiment was a 3x3 factorial design. The independent variables were recognition (B recognition, E recognition, and non-recognition) and training (B trained, E trained, and non-trained). The "training" variable refers to the sequential placement of critical word items in the five training sets. The "recognition" variable refers to the location of the critical word in a sixth, experimental CKT set. Eight female and seven male subjects were randomly assigned to each of the nine treatment conditions of the experiment.

Materials

Nine treatment conditions in the experiment were formed by manipulating the content of mock crime descriptions read by subjects prior to their polygraph examinations. A generic mock crime scenario is provided as an appendix at the end of this report. To manipulate the training variable, the descriptions given to B trained groups contained the following keywords: office, Rick, file cabinet, twice, and chest. Subjects in the E trained groups had these keywords: *hotel, Frank, desk, three,* and *head.* The non-trained subjects had the keywords: *bedroom, Kevin, dresser drawer, twice,* and *head* in their mock crimes. To manipulate the recognition variable, one third of the subjects in the B trained, E trained, and non-trained groups had either the additional keyword *red* in their crime (B recognition), the word *blue* (E recognition), or no color adjective (non-recognition).

Procedure

After completing a consent form and being given general information about the experiment, participants were presented with a pile of sealed envelopes and asked to select one. They were told to wait until the experimenter left the room and to then read the crime story inside the envelope. Each envelope contained one of nine versions of a description of a crime which involved a homicide and theft. These envelopes were prepared at the beginning of the experiment and shuffled thoroughly. Subjects were also instructed to answer "memory test" questions on the back of the crime scenario page, as proof that they had read it prior to their polygraph test. The experimenter waited in

an adjacent testing room for the subject to enter.

Participants were seated in a polygraph chair and the sensors were applied. Subjects' crime stories remained in the first room, so the experimenter did not know which version a subject had read. Before the test, subjects were verbally reminded that the purpose of the experiment was to determine whether the procedure was able to detect possession of crime-related information, and that they were to say the word "No" following each word item in six multiple choice sets. They were also reminded that they would be given a cash bonus of five dollars if they were found innocent on the test. After a three-minute baseline adjustment period, the recording of the CKT was played. The six CKT sets were as follows:

Where did the murder take place? Was it...
a, in a hospital room; b, in an office; c, in a bedroom; d, in a library; e, in a hotel room; f, in a living room?

What was the name of the person who was killed? Was it... a, Blain; b, Rick; c, Jim; d, Bill; e, Frank; f, Kevin?

Where did the killer find the gun? Was it...
a, on a bookshelf; b, in a file cabinet; c, in a closet; d, in a dresser drawer; e, on a desk; f, in a bar fridge?

How many times was the victim shot? Was it... a, six times; b, twice; c, four times; d, once; e, three times; f, five times?

What part of the victim's body was shot? Was it... a, his arm; b, his chest; c, his leg; d, his back; e, his head; f, his stomach?

What color was the special envelope? Was it... a, yellow; b, red; c, green; d, purple; e, blue; f, brown?

Unlike common practice, the letters preceding each word item in the test were spoken aloud on the tape recording. This was to help ensure that the subjects in the B trained and E trained groups would form an association between a specific letter (B or E) and critical words. Subjects in the non-trained groups were not expected to form such an association, since critical words occurred following different letters with equal frequency. All subjects heard the same recording.

After the test, the physiological sensors were removed and the subjects were directed back to the first room, where their polygraph charts were reviewed with them. They were then debriefed about the purposes of the experiment, paid and dismissed.

Electrodermal Response Quantification

Electrodermal responses were defined as any decrease in skin resistance with the initial inflection point falling inside a latency period of 0-5 s after onset of the word item, as indicated on the marker channel. Peak amplitude was measured up to 8 s after onset. Amplitude was measured in millimeters and then converted into standard deviate (z) scores following a procedure recommended by Ben-Shakhar (1985). Amplitude scores were converted relative to each subject's mean and standard deviation of responses to all 36 word items presented in the test. These z scores allow between-subject comparisons to be made on a common metric, even when there are substantial between-subjects differences in tonic level and phasic reactivity. In cases where there were unscorable responses, the z scores were calculated on the basis of the remaining responses. Of the 4680 stimulus presentations given to all subjects in the experiment, 126 (2.59%) individual responses were not scorable. Data omissions occurred with approximately equal frequency across experimental conditions.

Results

Experimental Set Response Magnitudes

Standardized skin resistance responses to item "B, *red*" in the sixth (experimental) CKT set were analyzed as the dependent variable in a 3x3 factorial analysis of variance (ANOVA), with training and recognition as independent variables. Cell means and standard deviations for the nine treatment groups are presented in Table 1. There was a significant main effect (F (2, 126)

6.27, p<.01) of recognition and a significant interaction (F (1,126) = 4.68, p<.01) between recognition and training.

To determine whether information recognition affects response magnitude, a planned contrast was carried out. It was found that the mean of the three B recognition groups (M=0.26, SD=0.80) was significantly greater (t (133) = 2.77, p<.01) than the mean of the six groups for whom *red* was not a critical word (M=-0.18, SD=0.92).

To determine whether violation of acquired expectancies affects response magnitude, a planned contrast was carried out. It was found that the combined mean of the E trained / B recognition and B trained / E recognition groups (M=0.58, SD=0.96) was significantly greater (t(133) = 4.35, p<.01) than the seven groups without expectancy violation (M=-0.20, SD=0.81).

The apparently low level of selective electrodermal response observed in the B trained / B recognition group (33.3%) led to a post hoc comparison which showed that the mean standardized electrodermal response (M=-0.10, SD=0.61) of this group was significantly smaller (t (43) = 2.23, p<.05) than the combined mean of the two other B recognition groups (M=0.44, SD=0.83).

Training Set Habituation Effects

To examine the possibility that responses to critical items may have habituated differentially relative to the control alternatives in the five training sets, Pearson correlation coefficients (r) between question set (numbered 1 through 5) and standardized skin resistance scores to critical and control words were calculated. The correlations between set number and responses to critical words for B trained, E trained, and non-trained groups were -.36, -.34, and -.30, respectively. The correlations between set number and responses to control words were -.25, -.38, and -.31, respectively. Each of these correlations differed significantly from a hypothetical correlation coefficient of r=0 (p<.01). None of the differences between observed correlations were significant at the .05 level.

Training Set Detection Rates

Rates of information detection were calculated for the five training sets using a procedure similar to that used by Lykken (1959). Electrodermal responses to the words following the letter B through F were ranked from greatest (1) through smallest (5). The first alternative in each set is generally considered to be a 'buffer' item, and was ignored for the purposes of classification.

Points were allocated on the basis of the rank of the critical item. If a rank of 1 was observed, two points were given; a rank of 2 earned one point, and no points were allocated for ranks of 3, 4, or 5. The points were tallied for all sets in which skin resistance was measurable for at least four of the alternatives, including the critical word. Across the three groups, Lykken scores were calculated on the basis of five sets in 109 cases, four sets in 22 cases, three sets in 3 cases, and one case had only 2 scorable sets. Using a cutoff point equal to the number of sets tallied for each subject, 76% of B trained, 58% of E trained, and 56% of non-trained subjects would be classified as informed. These differences in classification were not significantly different from one another (X^2 (2) = 5.20, p>.05). The B trained (X^2 (1) = 20.86, p<.001), E trained (X^2 (1) = 13.51, p<.001), and non-trained (X^2 (1) = 12.10, p<.001) detection rates were all significantly greater than a hypothetical chance hit rate of 20%.

Experimental Set Detection Rates

The number and percentage of subjects in each set who showed their largest electrodermal response in the sixth (experimental) CKT to item "B *red*" are shown in Table 1. To estimate the probability of each of these frequencies, a spreadsheet was generated which contained all possible combinations of ranks (1 through 5) that could be observed to one word item across 15 subjects. Of the 5^{15} combinations (over 30.5 billion), the number containing one rank of 1, two ranks of one, three ranks of one, and so on were computed, yielding a frequency distribution of hypothetical ranks. By dividing each of the cumulative frequencies of scores by 5^{15}, the chance likelihood of

obtaining each detection rate in a single sample of 15 subjects could be estimated. These probabilities are given as the p statistic in Table 1.

A one-tailed rejection region of .05 may be applied to this distribution by determining the smallest detection frequency with a cumulative proportion of hypothetical detection scores equal to or greater than that value which are smaller than .05. By doing this, we found that 6.105% of hypothetical detection scores were equal to or greater than six, but that 1.805% of hypothetical detection scores were equal to or greater than seven. It was therefore determined that groups in which seven or more subjects responded maximally to item "B, *red*" in the experimental set had a frequency of detection beyond what would be expected to occur by chance alone. Four such groups were identified: B trained / E recognition, B trained / non-recognition, E trained / B recognition, and non-trained / B recognition.

It is also possible to determine a one-tailed probability that a given observed detection rate is below what would be expected by chance. To do this, a similar procedure is employed, but the hypothetical cumulative frequency cutoff is set to .95. A total of 96.481% of hypothetical frequencies had one or more ranks of 1, whereas 100% of the hypothetical frequencies had zero or more. Therefore an observed frequency of zero ranks of 1 to "B *red*" would be significantly below what one would expect to occur by chance. Only one group was observed with detection below chance: E trained / non-recognition.

Experimental Set Signal Detection

Signal detection statistics are commonly used to estimate the psychometric efficiency of a test across all possible cutoff points (Bamber, 1975). Receiver Operating Characteristic (ROC) curves, and the area beneath those curves, were calculated for the B trained, E trained, and non-trained groups. ROC curves are generated by segregating responses of informed and non-informed subjects to critical items into two separate distributions, which are then ordered from largest magnitude to smallest. Cumulative percentages of responses falling above each observed level of response magnitude are then tabulated and plotted. The resulting curve has the likelihood of true positive outcomes along the y axis and the likelihood of false positive outcomes along the x axis. The area beneath this curve is an indicator of detection efficiency of the test across all possible cutoff points. An area of .50 represents chance, and positive and negative deviations from chance represent rates of correct classification above and below chance, respectively. Signal detection statistics were generated by comparing responses of B recognition and non-informed groups to item "B, *red*". ROC areas for the B trained, E trained, and non-trained groups were 0.57 (variance=.010), 0.81 (variance=.006), and 0.89 (variance =.003), respectively. ROC area and variance estimates were calculated by Gershon Ben-Shakhar of the Hebrew University of Jerusalem, using a special computer program.

To determine whether the ROC areas differed significantly from one another, the following formula was used to obtain probability estimates for the observed differences:

Z = (areal - area2) / square root (variancel + variance2)

Using this formula, normal deviate scores of 2.81 ($p<.05$), 1.89($p<.05$), and 0.81 ($p>.05$) were computed for the differences between E trained and B trained, non-trained and B trained, and E trained and non-trained groups, respectively. Only the first two of these contrasts indicated differences between informed and non-informed subjects that were significantly greater than what would be expected to occur by chance.

Table 1

Mean Standardized Electrodermal Responses, and Frequency with Largest Electrodermal Response to "B, *red*" in Experimental CKT Set, Across Nine Treatment Groups.

Group		Mean response, (SD)	Number detected	% detected	(p)
B trained					
	B recognition	-0.10 (0.62)	5	33.3%	-0.164
	E recognition	0.67 (1.06)	8	53.3%	(.004)*
	No recognition	-0.18 (0.83)	7	46.7%	(.01 8)*
E trained					
	B recognition	0.48 (0.88)	10	66.7%	(<.001)*
	E recognition	-0.53 (0.72)	1	6.7%	(.965)
	No recognition	-0.59 (0.48)	0	0.0%	(1.000)**
Non-trained					
	B recognition	0.40 (0.80)	7	46.7%	(.018)*
	E recognition	-0.15 (0.96)	4	26.7%	(.352)
	No recognition	-0.31 (0.91)	3	20.0%	(.602)

Legend: SD = standard deviation
* = significantly above chance detection
** = significantly below chance detection
p = prior probability

Discussion

Acquired expectancies can have powerful effects on the magnitude of physiological responses and on the likelihood of detection in subsequent CKT sets. The most interesting results of the present experiment were that subjects in the B trained / B recognition condition failed to show large electrodermal responses to the word "*red*" in the experimental set, despite the fact that they recognized the word as critical. Subjects in B recognition groups who were not conditioned to expect a critical word to follow the letter B in the experimental set displayed the large electrodermal responses to "*red*" that are typical of informed subjects tested with the CKT. Furthermore, large responses were observed in the B trained / E recognition and B trained / non-recognition groups, even though the word "*red*" was not a critical stimulus for those subjects. This somewhat peculiar pattern of physiological response resulted in true positive detection rates among B trained / B recognition subjects that were not significantly above chance, and rates of false positive detection among the B trained /E recognition and B trained / non-recognition groups that were significantly above chance. Also, the B trained condition produced an ROC area in the experimental set that was significantly lower than those of the E trained and non-trained conditions.

From these results, it would seem as though expectation of a critical word enables an informed subject to somehow avoid reacting to it, yet when an unexpected control word is presented, a physiological response occurs. This pattern lies in direct contrast to what one might anticipate if expectation had no effect. Subjects in the non-trained groups showed the more typical pattern, wherein electrodermal responses are elicited by critical words, and not by controls. The present results represent a direct refutation of Lykken's (1959) proposal that detection in the CKT is mediated solely by the recognition of critical information. That model can not account for the absence of differential responding among informed subjects to a critical item, as in the B trained B recognition condition. Nor does it explain informed subjects differential responding to a control item in the B trained/E recognition condition.

Information recognition may be sufficient to elicit a differential physiological response, but it is not necessary.

Implications For Theory of Information Detection

The notion that the CKT functions solely as an information detector is simply not supported by this evidence. To understand why selective physiological responses may come to be elicited by critical word items, a new theoretical perspective for the CKT must be developed.

At least two possible explanations for these results might be postulated. The first relates to Sokolov's (1963) notion of the orienting response (OR). Sokolov's theory of attention proposes that physiological responses are elicited by stimuli that are perceived as either significant (i.e. a critical word item) or novel (i.e. an unexpected word item). Yet, this account can not explain the lack of physiological response to the critical (significant) word seen in the B trained / B recognition subjects, without some modification to Sokolov's theory. The comparator mechanism would have to evaluate stimuli in terms of both novelty and significance, and somehow weight these two characteristics in order to determine whether the presented stimulus is above or below a certain threshold that is required for elicitation of the response. From such an information processing perspective, the most logical conclusion would have to be that the attention mechanism functions like a "gate" (Waters, McDonald, & Koresko, 1977) and that stimuli are attended to only if they are in some way important to the current functioning of the person. Either novelty or significance may allow the noteworthiness (Maltzman, 1977) of the stimulus to reach this attention threshold, but if a significant stimulus is expected, then the threshold is raised. Factors such as guilt, motivation to avoid detection, and deceptive verbal responses may also contribute to the likelihood of detection by adding significance to the critical items. According to this model, information recognition holds no special status, other than as one of several contributors to stimulus relevance, which may also be affected by motivation, guilt, and deception. This account bears some similarity to the attention-based theories of information detection proposed by Elaad & Ben-Shakhar (1 989) and by Waid, Orne, Cook, & Orne (1978). It also has some similarity to the revised dichotomization models (e.g. Ben-Shakhar & Gati, 1987; Gati, Ben-Shakhar, & Avni-Liberty, 1996), which postulate that novelty and significance factors jointly contribute to OR.

A second explanation might be proposed from a motivational perspective. Although the circumstances surrounding the polygraph examinations in the present experiment made them have far less gravity than is typical outside the laboratory, the instructions given to subjects were intended to motivate them to attempt to 'pass' the test and appear innocent. Because of this motivation, one might expect subjects' reactions to critical word items to be determined by at least two competing response tendencies. Recognition of a word as critical might dispose a subject toward autonomic reaction. On the other hand, since the subject is motivated to attempt to appear innocent, there may be motivation to attempt to suppress this arousal. This situation is analogous to an explanation of Unconditioned Response Diminution (the tendency for response to a stimulus to decrease with repeated presentations (Goddard, 1991)) known as opponent process theory (e.g. Schull, 1979). When applied to the CKT, this theory predicts that the magnitude of a response should consist of two opposing forces: a tendency for momentarily increased autonomic activation following critical word items, and a tendency for the subject to attempt to suppress this arousal. Such a model does account for selective responding of informed subjects to critical items who are not prepared for the presentation of those words. It also explains why subjects in the B trained / B recognition condition of the present experiment failed to show selective responses to the critical word, because their tendency to respond was countered by whatever compensatory maneuvers they may have marshaled in order to decrease the response. It also explains why subjects in the B trained / E recognition condition reacted

strongly to "*red*", which was a non-critical word for them. They were prepared for a critical word, but they had no preparation for a control. The B trained / non-recognition subjects may have reacted to "*red*" because they inferred it to be a critical word, and this additional process may have resulted in a momentary increase in arousal.

This latter theory has some interesting implications for an issue well known to all forensic psychophysiologists, as well as to a disturbing number of examinees: the problem of countermeasures. If countermeasures can become associated with a suspect's compensatory response to critical items, or as a tendency to augment responses to controls, one way to foil such maneuvers might be to employ some form of expectancy manipulation. To illustrate, imagine a CKT consisting of several sets containing information that the suspect is known to be aware of, followed by a discriminative set consisting of a fact that would be known only by the examiner and a guilty suspect. A clever suspect might try to augment their responses to control words and to ameliorate their responses to the critical words, In the training sets, they would learn when to do this. Suppose that the test had the same structure as the one used in the present experiment (which may or may not be optimal) and the critical words follow E in the training sets. The suspect would learn to attenuate responses to words following E and to augment responses to words following B, C, D, and F. If a subsequent discriminative set had the critical item at a position other than E, the informed (guilty) suspect would be ill-prepared to combat their tendency to react. Indeed, if they applied the tactic of enhancing responses to controls, they might make the mistake of augmenting their reaction to the unexpected critical word! On the other hand, an uninformed (innocent) suspect would not realize that anything is awry, infer that whatever word follows E is critical and respond maximally to it, thereby unintentionally reducing the likelihood that they could be falsely incriminated. These possibilities should be explored experimentally and not necessarily with respect only to the CKT; similar anti-countermeasure tactics using expectancy manipulation might also be developed in the context of other tests, such as the POT and CQT.

Implications For Field Application of CKT

Although useful as an experimental paradigm, the CKT is inherently difficult to apply in actual investigative settings. In order to reduce the likelihood of false positive error, an examiner must create a test that consists of several multiple-choice sets, each containing one critical item and several control alternatives. Such a strategy greatly adds to the difficulty in constructing the test. Correct answers must be known by the guilty suspect, unknown by innocent suspects, and be available to the examiner when devising the test questions. To prevent false incrimination of innocent suspects, the information used in the test must be absolutely secure from leakage via media reports, hearsay, etc. That same relevant information also must be stored in the memory of the guilty suspect in order for it to be recognized at the time of the test. Trivial bits of information that are typically found at a crime scene by forensic investigators might not be encoded and stored in the suspect's memory and would therefore be poor candidates for CKT items. Moreover, the relevant information must be of a type that is amenable to being put in a multiple choice format with control alternatives that are similar to it and of equivalent arousal value. Despite the CKT's attractive rationale, practical problems such as these limit its application. Elaad & Ben-Shakhar (1997) recently addressed this problem by demonstrating impressive rates of correct classification using a single CKT set repeated many times. Although this partially solves the problem of application of the CKT when suitable information is limited, the logistical obstacles are often so great that creating even one CKT set is difficult.

There sometimes are cases when information suitable for use in a CKT is available, but in very limited quantity. Under such circumstances, it might be possible to use either Elaad & Ben Shakhar's (1997) repeated-item CKT, or a modified version which employs expectancy manipulation. Detection rates among the E trained conditions of this experiment were

encouraging, with two thirds of B recognition subjects correctly detected and no false positives among non-informed subjects. Both of these results were significantly beyond chance. Although a 66.7% hit rate may not sound overly impressive, it was derived from a single CKT set, under conditions of modest external validity. Also, detection estimates in this study may have been affected by the fact that subjects merely read information about a mock crime, but did not enact it. This is similar to the "informed innocent" conditions of Bradley & Warfield (1984) and Bradley & Rettinger (1992), which produced lower rates of detection compared to their 'mock crime guilty" subjects, who enacted the mock crimes. Still, detection in the E trained condition equaled that of Elaad & Ben-Shakhar's repeated-item CKT, which correctly identified 67% of mock crime guilty subjects in their high motivation condition.

While it is not recommended that any test use a single response to discriminate guilt from innocence, the expectancy effect might be robust enough to provide some improvement upon Elaad & Ben-Shakhar's repeated-item CKT. Such a modified test might consist of a series of training sets containing disclosed information, followed by a number of repetitions of an undisclosed discriminative set. Each repetition might have the critical alternative in a location that does not conform to the expectation acquired through conditioning. It is possible that the association between specific letters and critical words might extinguish after several presentations that violate the association, but it is also possible that the rate of detection with a single item of undisclosed information could be enhanced by using such an amalgam of the repeated-item and expectancy manipulation techniques. Further research along these lines might result in an improved form of CKT that is both effective and applicable in a wider range of cases.

If some form of expectancy manipulated CKT were ever to be applied as anything other than a research instrument, we would strongly recommend that the training sets not be used to infer guilt or innocence. Whether or not a guilty suspect might react more strongly to the critical items in the training sets than would an informed innocent suspect, the use of disclosed information to infer culpability would be a very dangerous practice. It would also be advisable to ensure t at every suspect possesses the information used in the training sets, and this should be verified prior to the test. The whole point of this procedure is that both guilty and innocent subjects be privy to the information used in the training sets, and any scoring criterion which incorporates those sets would almost certainly be prone to false positive error. Those sets should be used only to manipulate expectations about the presentation of critical information in the subsequent discriminative set(s). Detection rates and habituation results for the training sets were reported here solely for the purpose of future comparison with other studies.

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. (1985). Standardization within individuals: A simple method to neutralize individual differences in psychophysiological responsivity. *Psychophysiology,* 22, 292-299.

Ben-Shakhar, G., & Dolev, K. (1996). Psychophysiological detection through the guilty knowledge technique: Effects of mental countermeasures. *Journal of Applied Psychology,* 81, 273-281.

Ben-Shakhar, G., & Furedy, J. J. (1990). *Theories and Applications in the Detection of Deception.* New York: Springer-Verlag.

Ben-Shakhar, G., & Gati, I. (1987). Common and distinctive features of verbal and pictorial stimuli as determinants of psychophysiological responsivity. *Journal of Experimental Psychology: General,* 111, 91-105.

Bradley, M.T., & Janisse, M.P. (1981). Accuracy demonstrations, threat and the detection of deception: cardiovascular, electrodermal and pupillary measures. *Psychophysiology,* 18, 307-315.

Bradley, M.T., MacLaren, V.V., & Carle, S.B. (1996). Deception and non-deception in guilty knowledge and guilty actions polygraph tests. *Journal of Applied Psychology,* 81, 153-160.

Bradley, M.T., & Rettinger, J. (1992). Awareness of crime-relevant information and the guilty knowledge test. *Journal of Applied Psychology,* 77, 55-59.

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

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

Davis, R.C. (1961). *Physiological Responses as a Means of Evaluating Information.* in A.D. Biderman and H. Zimmer (eds.), The Manipulation of Human Behaviour. New York: Wiley. pp. 142-168.

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

Elaad, E. (1993). The role of guessing and verbal response type in psychophysiological detection of concealed information. *The Journal of Psychology,* 127, 455-464.

Elaad, E., & Ben-Shakhar, G. (1989). Effects of motivation level and verbal response type on psychophysiological detection in the guilty knowledge test. *Psychophysiology,* 26, 442-451.

Elaad, E., & Ben-Shakhar, G. (1991). Effects of mental countermeasures on psychophysiological detection in the guilty knowledge test. International Journal of Psychophysiology, 11, 99-108.

Elaad, E., & Ben-Shakhar, G. (1997). Effects of item repetitions and variations on the efficiency of the guilty knowledge test. *Psychophysiology,* 34, 587-596.

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology,* 77, 757-767.

Furedy, J.J., & Ben-Shakhar, G. (1991). The role of deception, intention to deceive, and motivation to avoid detection in the psychophysiological detection of guilty knowledge. *Psychophysiology,* 28, 163-171.

Gati, I., Ben-Shakhar, G., & Avni-Liberty, S. (1996). Stimulus novelty and significance in electrodermal orienting responses: The effects of adding versus deleting stimulus components. *Psychophysiology,* 33, 637-643.

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, 111.

Goddard, M. (1991). US-US associations as a factor in the habituation to emotionally arousing stimuli. *Motivation and Emotion,* 15, 207-219.

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

Gustafson, L.A., & Orne, M.T. (1965). The effects of verbal responses on the laboratory detection of deception. *Journal of Applied Psychology,* 49, 412-417.

Honts, C.R., Devitt, M.K., Winbush, M., & Kircher, J.C. (1996). Mental and physical countermeasures reduce the accuracy of the concealed knowledge test. *Psychophysiology,* 33, 84-92.

Horneman, C.J., & O'Gorman, J.G. (1985). Detectability in the card test as a function of the subject's verbal responses. *Psychophysiology,* 22, 330-333.

Iacono, W.G., Cerri, A.M., Patrick, C.J., & Fleming, J.A.E. (1992). Use of antianxiety drugs as countermeasures in the detection of guilty knowledge. *Journal of Applied Psychology,* 77, 60-64.

Kugelmass, S., & Lieblich, 1. (1966). The effects of realistic stress and procedural interference in experimental lie detection. *Journal of Applied Psychology,* 50, 211-216.

Kugelmass, S., Lieblich, I., & Bergman, Z. (1967). The role of "lying" in psychophysiological detection. *Psychophysiology,* 3, 312-315.

Lieblich, I., Naftali, G., Shmueli, J., & Kugelmass, S. (1974). Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. *Journal of Applied Psychology,* 59, 113-115.

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 effects of faking. *Journal of Applied Psychology,* 29, 725-739.

Maltzman, I. (1977). Orienting in classical conditioning and generalization of the galvanic skin response to words: An overview. *Journal of Experimental Psychology: General,* 106, 111-119.

Schull, J. (1979). A conditioned opponent theory of Pavlovian conditioning and habituation. *The Psychology of Learning and Motivation,* 13, 57-90.

Sokolov, E. N. (1963). *Perception and the Conditioned Reflex.* New York: MacMillan.

Stern, R.M., Breen, J.P., Watanabe, T., & Perry, B.S. (1981). Effects of feedback of physiological information on responses to innocent associations and guilty knowledge. *Journal of Applied Psychology,* 66, 677-681.

Waid, W.M., Orne, E.C., Cook, M.R., & Orne, M.T. (1978). Effects of attention, as indexed by subsequent memory, on electrodermal detection of information. *Journal of Applied Psychology,* 63, 728-733.

Waters, W. F., McDonald, D.G., & Koresko, R. L (1977). Habituation of the orienting response: A gating mechanism subserving selective attention. *Psychophysiology,* 14, 228-236.

Appendix

Sample Instruction Sheet

Instructions

Please read these instructions twice

You are a suspect in a brutal murder. You did not do it. You are not even capable of doing it. However, you have no witnesses to account for what you were doing on the day of the crime, and the interrogations with the police have not been going well. From these interrogations and the newspaper you have learned quite a bit about the crime.

A fairly big time crook name (NAME) was murdered in (LOCATION). He was shot (NUMBER) times in the (BODY PART). The gun was his own and had been taken from a (LOCATION IN ROOM) and then after it had been used was thrown in a waste basket. To the police, this indicated the work of an amateur because a professional would have his own weapon. Rick was not murdered for money since he had only one dollar in his pocket. It seems he was murdered for information he carried in a (COLOUR) envelope.

The police have accused you of the murder, and you don't have any proof that you are innocent. However, there is one thing you can do to strengthen your case before the prosecution goes any further. You are going to be given a polygraph lie detector test. If you are found innocent on the test, then you might not even have to go to court. But, if you are found guilty, then you may face a very grim future.

The test is kind of like a multiple choice exam. You will hear questions like, "After you killed the man, where did you hide the gun? Was it... a, in a wastebasket; b, in a locker; c, in the trunk of a car... etc.". After each alternative, you will say "No.". That means that when you hear the correct answer, you will be lying if you are guilty. The examiner will monitor your physiological responses each time you say "No.". If you appear to be lying by saying "No." to the right answers, then you will be found guilty.

If you are found innocent on the test, you will be given a cash bonus of five dollars. If you are found guilty, you will receive nothing.

Fiscal Year 1997 Report to Congress on the Department of Defense Polygraph Program

Executive Summary

The Department of Defense (DOD) uses the polygraph in criminal investigations, counterintelligence cases, foreign intelligence and counterintelligence operations, and exculpation requests. This report contains numerous examples of polygraph utility in resolving counterintelligence and security issues as well as criminal investigations. The polygraph is clearly one of our most effective investigative tools.

About 67 percent of our polygraph examinations are conducted as a condition for access to certain positions or information under the DOD Counterintelligence-Scope Polygraph (CSP) program. The purpose of the CSP Program is to deter and detect activity involving espionage, sabotage, and terrorism. In Fiscal Year 1997, the Department implemented changes to the CSP Program to reduce the intrusiveness of polygraph screening examinations while providing maximum standardization and ensuring reciprocity within the Intelligence Community. The Department also implemented some new initiatives increasing the continuing education requirement for polygraph examiners, providing a quality control assurance program, expanding our information database and increasing our use of computer-based and off-site training to reduce travel costs.

The Department conducts CSP examinations on military personnel, DOD civilian employees, and DOD contractor personnel. Of the 7,616 individuals examined under the CSP Program in Fiscal Year 1997 7,440 showed no significant physiological response to the relevant questions (non-deceptive) and provided no substantive information. The remaining 176 individuals yielded significant physiological responses, or were evaluated as inconclusive and/or provided substantive information. Of these 176 individuals, 154 received a favorable adjudication, two are still pending adjudication, 14 are pending investigation, and six individuals received adverse action denying or withholding access.

The Department of Defense Polygraph Institute (DODPI) trains all federal polygraph examiners. The basic polygraph courses are taught at the Masters Degree level. The Institute also offers specialized courses in forensic psychophysiology through their continuing education program. In addition, the Institute conducts on-going evaluations of the validity of polygraph techniques used by the Department as well as research on new polygraph techniques, instrumentation, analytic methods, and polygraph countermeasures. The DOD research program is authorized by Public Law 100-180.

I. DOD Use of Polygraph Examinations

The Department of Defense has used the polygraph for almost half a century. It is used in criminal investigations, counter-intelligence cases, foreign intelligence and counterintelligence operations, exculpation requests, and as a condition for access to certain positions or information. The polygraph is a tool that enhances the interview and interrogation process. Often it is the only investigative technique capable of providing essential information to resolve national security issues and criminal investigations. The use of the polygraph as a condition for access is limited by a statutory quota for Counterintelligence-Scope Polygraph (CSP) examinations.

The following table reflects Department of Defense Polygraph Program statistics for fiscal year 1997:

Criminal	2,338	20.6%
Exculpatory	565	5.0%
CI Scope	7,616	67.2%
All Others*	812	7.2%
Total**	11,331	100%

* Includes examinations conducted in support of personnel security investigations, counterintelligence and intelligence operations, and polygraph assistance to non-DOD federal agencies.

** Does not include polygraph examinations conducted by the National Security Agency (NSA). A breakout of polygraph examinations conducted by NSA is contained in a classified table submitted with this report. Nor does it include polygraph examinations conducted by the National Reconnaissance Office, which are conducted under the authority of the Director of Central Intelligence (DCI).

II. Fiscal Year 1997 Counterintelligence-Scope Polygraph Examinations

Section 1121 of the National Defense Authorization Act for Fiscal Years 1988 and 1989 (Public Law 100-180, December 4, 1987; 101 Stat. at 1147) authorizes the Department of Defense to conduct Counterintelligence-Scope Polygraph (CSP) examinations as a condition for access to certain information.

The purpose of the CSP Program is to deter and detect espionage, sabotage, and terrorism. The following topics are covered during the CSP examination: (1) Involvement with a foreign intelligence/security service, involvement in espionage; (2) Involvement in terrorism; (3) Unauthorized foreign contacts; (4) Deliberate failure to protect classified information; and (5) Damaging/sabotaging government information systems, clandestine collection, or defense systems. These CSP topics meet the needs of both DOD and the Intelligence Community facilitating the transfer of security clearances.

In Fiscal Year 1997, the Department modified the procedures for conducting CSP examinations to reduce the intrusiveness of CSP examinations, increase their standardization, and maximize reciprocity within the Intelligence Community. Also, there is increased emphasis on aperiodic, rather than periodic, examinations, which provide a greater deterrent. In addition, the Department has implemented new initiatives increasing the continuing education requirements for polygraph examiners, providing a quality control assurance program, expanding our information database, and increasing our use of computer-based and off-site training to reduce travel costs.

Public Law 100-180 authorizes DOD to administer CSP examinations to persons whose duties involve access to information that has been classified at the level of top secret or designated as being within a special access program under section 4.2(a) of Executive Order 12356 (superseded by Executive Order 12958). This includes military and civilian personnel of the Department and personnel of defense contractors. The number of CSP examinations has been limited to 5,000 per fiscal year since Fiscal Year 1991. During Fiscal Years 1988 through 1990 the ceiling was 10,000. The quota reduction took place two years after new exemptions for cryptographic and reconnaissance programs were adopted. Public Law 100-180 exempts certain intelligence agencies and functions from the 5,000 quota: (1) individuals assigned, detailed or under contract with the Central Intelligence Agency (CIA), (2) persons employed, assigned, detailed, under contract or applying for a position in the National Security Agency, (3) persons assigned to a space where sensitive cryptographic information is produced, processed, or stored, and (4) persons employed by, assigned or detailed to, an office within the Department of Defense for the collection of specialized national foreign intelligence through reconnaissance programs or a contractor of such an office.

The following table reflects CSP examinations conducted by the Department of Defense in accordance with Public Law 100-180.

(1) Special Access Programs	1,670
(2) DIA Critical Intelligence Positions	994
(3) TOP SECRET	0

Question Formulation

Norman Ansley

Abstract

This paper contains observations about question formulation for polygraph testing followed by specific guidelines. Applicable to the frequently used testing formats, the guidelines cover relevant questions, probable lie control/comparison questions, irrelevant questions, and wording of peak of tension and guilty knowledge tests. The paper does not offer guidelines for technical questions used by only one test format. There are abstracts of three legal cases where question formulation was an issue. The references cited are included in a larger reference section.

Keywords: Control question, comparison question, guilty knowledge test, irrelevant question, peak of tension, question formulation, relevant question, semantics.

General Observations

One of the clinical aspects of polygraph testing is the formulation of questions. Some guidelines are suggested in this paper.

A word of caution about prepared question lists and notepacks. Blind adherence may result in the examinee not understanding one or more questions, causing problems in testing. Written questions are a good guide to policy, but the suggested words may not be in the examinee's vocabulary. The advantage of prepared lists and notepacks is better compliance with policy, regulations and law.

When working with investigators or attorneys who do not understand the limits of polygraph testing, you may be presented with a lengthy list of poorly worded questions that you cannot use. Ask them to describe the case and suggest one question, answered "yes" or "no" that will solve it. Try to conduct your tests with a single-issue test format, as they tend to be more accurate than multiple issue formats. More issues create more opportunities for error.

Be wary of precisely worded relevant questions proffered by the examinee or his attorney. The question may avoid the issue or be part of an effort to rationalize.

An examinee will not readily admit he does not understand a question. The lack of understanding shows up when the examinee is asked to explain why the question is being asked and what it means.

When the questions are agreed upon, and they exclude details or the wording is a bit unusual, be sure the missing details and a discussion of the development of the relevant questions are in the report. Details that were agreed upon, but were deleted from the question, must be in the report. Persons who were not present may criticize the relevant question wording because the report does not adequately describe the question development.

In screening applicants, keep in mind that EEOC and ADA rules on job interviewing apply to polygraph testing. For example, under ADA you cannot ask medical questions until a bona fide offer of employment is made,

The author is a Life Member of the APA and president of Forensic Research, Inc. For reprints write to him at 35 Cedar Road, Severna Park, Maryland 21146.

To make this paper into a workbook for teaching, 35 questions may be obtained from the author. Each of the questions has one or more errors in wording. The student identifies the errors, then writes a correct version of the questions.

and the questions you usually ask to determine fitness for testing are considered medical, you either don't ask, or have the polygraph tests performed after the offer stage. In addition to the Federal limits, there are state laws and city ordinances that further limit what you can say.

The technical questions that are designed to appear as relevant questions must be treated with the same thoroughness as the relevant questions. Included are the control/comparison questions (except in PCQT and DLC), sacrifice relevants, and the identity irrelevants in some RI tests.

While keeping a question short is often desirable for clarity, it is not essential. I have seen long and complex questions used in contract fraud, and the tests were successful.

Some technical questions such as the sacrifice relevant (Capps, 1991; Horvath, 1994) and the symptomatic (Capps, Knill & Evans, 1993) have been the topic of specific papers. Much has been written about techniques and questions for disclosure and maintenance tests in sex offender tests. It is too early to comment on those questions or suggest guidelines. Wording of relevant and control/comparison questions in certain types of crimes suggest the need for expert advice. Examples are arson, bomb cases, contract fraud, and insider trading of stocks or commodities.

Relevant Questions - Guidelines

The relevant question must solve a vital problem.

The issue covered by the relevant question must be of vital importance to the examinee.

The question must pose a dichotomy, answerable by "yes" or "no."

The question must be fully understood and mean the same thing to the examiner and examinee.

When possible, a relevant question should not use legal or technical terms.

The question must not contain obscene, profane, racial, derogatory, degrading, or insulting words or phrases.

Qualifiers, such as "Other than ..." are placed at the beginning of the question.

There should be enough facts in the question to avoid outside issues.

There should be no more facts in the question than necessary.

The facts in the question should not only be correct, but would be recognized as correct by the perpetrator.

The question must not imply or assume guilt.

The question must not imply disbelief by the examiner.

The sentence must be a question (POT/GKT exception).

It is preferable to use the action (verb) rather than the result.

The question must not ask for an opinion.

The question should not give away facts you plan to use in a POT/GKT.

It is generally held that you cannot test on the issue of intent.

When testing victims, the issue is truthfulness, not rape, robbery, or some other crime.

Be wary of using specific amounts of money stolen in the question.

Avoid words that are emotional, and likely to cause a response.

Separate relevants are asked about direct involvement, secondary involvement, guilty knowledge, and evidence connecting facts.

Control/Comparison Questions

The final written descriptions of control/comparison questions are by Summers (1939) whose "emotional standards" questions were paired with relevant questions. Examples he gave included, "Were you ever arrested?" "Are you living with your wife?" and "Do you own a revolver?" From the text and examples it appears that Summers used probable lies, embarrassing, evidence connecting, and other questions. Inbau & Reid (1948) introduced a test format in the 1940s which included a probable lie and a guilt complex question for comparison purposes. The guilt complex was later dropped for a second probable lie. The Reid control question may include the offense at issue. However, the Backster (1969) technique and DoDPI control/comparison question techniques do not permit the control/comparison questions to include the offense. They separate relevant and control/ comparison coverage and offense by date or location.

The guilt complex question is a known-truth answer to what appears to be a relevant question about a crime. Other comparison questions include the yes answer to the relevant question in the PCQT format, a directed lie to a trivial matter, and the situational control where the examinee confirms and inculpatory fact with a yes answer. In one relevant-irrelevant screening test format, a relevant question with a low base rate of deception, such as terrorism, may serve as a probable truth (guilt complex) comparison question.

Probable Lie Control/Comparison Question Guidelines

The control/comparison question must be treated as a relevant question.

It is broader in scope than a relevant in order to be more likely applicable.

It is usually on the same topic as the case issue, but slightly lesser in severity of offense.

It should not mention or imply sex, except where sexual behavior is the issue.

Qualifiers such as "OT" from admissions should be at the beginning of the question.

Time bars should be used or not used, depending on the rules for the format.

It is usually worded to be answered "No."

When possible it should use the same verb as is in the relevant question.

The topic of the question should be one the examinee is likely to lie about or have serious doubt regarding the truthfulness of the reply.

Do not use a control/comparison question on race, religion, or politics, or that will humiliate or embarrass the examinee.

The question must be fully discussed with the examinee.

Irrelevant Questions

Almost all test formats open with an irrelevant question. Some formats anticipate additional need for an irrelevant question and fix its place in the format, while other techniques allow the examiner to insert them as needed. Irrelevant questions allow the orienting and other reactions to return to baseline, establish a norm level, reduce general nervous tension, provide relief from a previous reaction, separate reactions to relevant questions, and confirm the identity of the person being tested.

There are two types of irrelevant questions. One is the obvious irrelevant question, such as "Are you wearing brown shoes?" The other type of irrelevant involves identity questions, and is disguised as a relevant. These involve name, date and place of birth, residence, etc. Both of these types of irrelevant questions have a place in testing, and the type is sometimes prescribed. However, favoring the identity questions, Weir (1974) notes that the obvious irrelevants appear ridiculous, seem like a game, and do not pose a threat to the examinee.

Regarding research, Kircher and Raskin (1986) found that examinees were aware that the irrelevant questions produced their weakest reactions; and Frisby (1979) found that identity irrelevants produced fewer responses than did obvious irrelevants.

Irrelevant Questions - Guidelines

Identifying irrelevants are treated as relevants and thoroughly discussed.

Consistent significant reactions to identify irrelevants warrant interrogation.

Irrelevants must pose a dichotomy, answerable by "yes" or "no."

Irrelevants, obvious or identity, must be discussed with the examinee.

Answered truthfully, an irrelevant should not provoke emotions.

The proposed irrelevant is not a question you expect an examinee to lie to.

An obvious irrelevant is not related to the topic at issue.

Irrelevant questions are usually worded to be answered "yes." However, the Marcy and Arther CQT formats have obvious irrelevants answered "no." Several irrelevants must be reviewed before the test if the examiner is allowed by the technique to insert irrelevants as needed.

Most test formats open with one irrelevant, and some open with two.

Peak of Tension and Guilty Knowledge Tests

The peak of tension group includes the known solution peak (Type A) in which the investigator and the perpetrator know some specific item of information which would not be known to someone who was not involved in the offense. There is a variant called the guilty knowledge test (GKT). The primary difference between the POT and GKT is that in the latter the key item is placed by chance in the list anywhere except the first position, while in the POT the key is in or near the middle. There is a searching peak of tension (SPOT) in which the examiner is seeking to locate evidence from a subject who may possess information he refuses to divulge, such as the location of loot, or location of the victim of a kidnapping. The stim or acquaintance test is in the peak of tension group. There are many variations of the stim, with a number described in a special issue of *Polygraph* (1978) 7(3) 173-215. Stim tests differ from most POT formats in that the series is asked only once, where most POT, GKT, and SPOT tests employ three series, often varying the sequence in each presentation.

Known Solution (Type A) and Guilty Knowledge Test - Guidelines

Place the key item in or near the middle of the list. In the GKT the key is to be placed by chance, but not at the beginning.

Be certain the key is the correct item.

Be certain that other items in the list cannot possibly be correct.

Be certain the guilty or involved would recognize the correct item.

Be certain the innocent would not know the correct item.

Be certain that concealing recognition of the key is important.

Try to keep all the items of similar length, one word, two words, etc.

Try to keep all the items of similar emotional content.

Do not include an absurd or illogical item.

You may use a logical sequence to the items, if the key is not first.

Five, six, or seven items are ideal, but more may be used if logical.

The examinee may be given the order, or a list posted.

To avoid dissociation have the examinee repeat the item before saying "No."

If you plan to give a POT after an RI or CQT, be sure the key item(s) are not given away in the questions or pretest.

All items must be discussed with the examinee.

If the list includes guns or cars, be certain the examinee is sufficiently knowledgeable to recognize calibers, makes, and models.

If you use Arther's false key, place it at number 2, and the key at 4 or later.

Searching Peak of Tension Tests (Type B, SPOT)

The most probable item should be in the middle of the list during the first of three presentations.

The least probable item should be at the beginning of the list during the first of three presentations.

Use a question about other possibilities as the last item on each chart.

The order should be varied with each presentation.

The order of items may be announced or posted.

Concealing the correct item must pose an obvious threat.

The items in the list should be discussed in detail.

When maps or diagrams are used, they must have clearly marked boundaries, numbers, letters, and names for each area.

Question Formulation - Legal Opinions

In *United States v. Lech*, 94 Cr. 285, 895 F.Supp. 582 (USDC SD NY 1995), a bribery case before federal trial Judge Sonia Sotomayer, defendant Wlodek Jan Lech, attempted to enter into evidence the results of a polygraph examination in which he answered such questions as, "Did you try to bribe any Board of Education officials to obtain asbestos removal contract?" and "Did you take part in trying to bribe Board of Education officials to obtain an asbestos removal contract?" Lech sought admissibility in light of *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 113 S.Ct. 2786. Judge Sotomayer did not address *Daubert* and Rule 702. She applied Rule 403 of the Federal Rules of Evidence and found Lech's polygraph evidence precluded because "its probative value is substantially outweighed by the danger of unfair prejudice, confusion, or misleading of the jury." She explained that "Each of the questions Lech seeks to introduce calls for his belief about the legal implications of his actions, without setting forth the factual circumstances underlying such conclusion." In other words, she wrote, "the jury would receive evidence showing Lech's personal belief that he did not violate any federal criminal statute, but would not receive any information that would assist its inquiry to find facts." In a footnote, the Judge indicated the outcome may be different if a defendant sought to introduce answers "to an exam where he or she completely denied any connection or involvement" with the alleged crime. [*New York Law Journal*, 28 July 1995]

In *Hester v. Milledgeville*, 777 F.2d 1492 (11th Cir. 1985) the Eleventh Circuit overruled a trial court's conclusion that the use of control questions was a violation of the Constitutional right to privacy. The appellate court said the City's interest in using control questions to improve the accuracy of the polygraph test is an important one ... and the specific control questions at issue constituted only a limited intrusion into the sphere of confidentiality. The Court noted that the questions were general in nature, were asked for a specific, limited purpose, and, although potentially embarrassing, avoided issues such as those related to marriage, family and sexual relations generally considered to be the most personal. The Eleventh Circuit issued a word of caution, saying they would have reservations if any governmental unit were to use a subject's response to a control question for any purpose other than comparing the polygraph reading for the control question to the same subject's reaction to a relevant

question. The Court added there might well be a point at which a control question is so embarrassing or specific, or concerns so personal a matter, as to render the question unconstitutional even when asked for the proper purpose.

In *State v. Stowers*, 580 S.W.2d 516 (Mo.App. 1979) the defendant was appealing conviction for forcible rape. The results of a stipulated polygraph examination had been admitted, and on appeal defendant said one of the questions asked during the test was factually inaccurate. The question at issue was "Did you rape ... on Route FF?" Defense stated that the prosecutrix testified that the rape was along a gravel road just off Route FF, and that inaccuracy should cast doubt over the reliability of the whole test, thus rendering it inadmissible. The Missouri Court of Appeals said that the reference to geographic area was sufficiently proximate to the crime site not to invalidate the test results.

References

Abrams, S. *The complete polygraph handbook.* Lexington, MA: Lexington Books, 1989.

Arther, R.O. (1982). How to word peak questions. *Journal of Polygraph Science*, 17 (2), 1-4.

Arther, R.O. (1987). Irrelevant questions. *Journal of Polygraph Science*, 22 (2), 1-4.

Backster, C. (1996). The Backster zone comparison technique, an update from the source. Paper presented at the American Polygraph Association Seminar, New Orleans, LA.

Backster, C. (1974). Reducing inconclusive results by pre-testing relevant questions. *Polygraph*, 3 (2), 216-222.

Backster, C. (1969). Standardized polygraph examination notepack, 5th ed. rev. San Diego, CA: Backster Associates.

Capps, M.H. (1991). Predictive value of the sacrifice relevant. *Polygraph*, 20 (1), 1-6.

Capps, M.H., Knill, B.L., & Evans, R.K. (1993). Effectiveness of the symptomatic questions. *Polygraph*, 22 (4), 285-298.

Department of Defense Polygraph Institute (1989). *Modified general question technique summary.*

Department of Defense Polygraph Institute (1991). *Technical correction of Curriculum* [regarding the peak of tension test.]

Department of Defense Polygraph Institute (1993). *Test question construction.*

Frisby, B.R. (1979). Research into the semantics of the irrelevant question. *Polygraph*, 8 (3), 205-223.

Harman, G.W. & Reid, J.E. (1982). The selection and phrasing of lie-detector test control questions. *Polygraph*, 11 (1), 82-86.

Harrelson, L.H. (1964). *Keeler Polygraph Institute training guide.* Chicago, IL: The Keeler Polygraph Institute.

Horvath, F. (1991). The utility of control questions and the effects of two control question tests in field polygraph techniques. *Polygraph*, 20 (1), 7-25.

Horvath, F. (1994). The value and effectiveness of the sacrifice relevant question: An empirical assessment. *Polygraph*, 23 (4), 261-279.

Inbau, F.E. (1942). *Lie detection and criminal interrogation.* Baltimore, MD: Williams & Wilkins.

Inbau, F.E. (1948). *Lie detection and criminal interrogation*, second ed., rev. Baltimore, MD: Williams & Wilkins.

Inbau, F.E. & Reid, J.E. (1953). *Lie detection and criminal interrogation*, third ed., rev. Baltimore, MD: Williams & Wilkins.

Kircher, J.C. & Raskin, D.C. (1986). Visceral perception in the detection of deception. SPR Abstracts, *Psychophysiology*, 23 (4), 446.

Koppang, C.E. (1985). *Polygraph examination student manual.* Ottawa: Canadian Police College.

Lee, C.D. (1953). *The instrumental detection of deception, the lie test.* Springfield, IL: Charles C Thomas.

Matte, J.A. (1996). *Forensic psychophysiology using the polygraph.* Williamsville, NY: J.A.M. Publications.

Minor, P.K. (1989). "The relevant-irrelevant technique," chapter 10 in S. Abrams, *The complete polygraph handbook.* Lexington, MA: Lexington Books.

Reid, J.E. & Inbau, F.E. (1966). *Truth and deception: The polygraph (lie detector) technique.* Baltimore, MD: Williams & Wilkins.

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

Smith, M.L. (Moderator). (July 1993). Polygraph test question formulation ..." A panel discussion. Handout at the Seminar of the American Polygraph Association, Newport, CA.

Special issue on stimulation tests (1978). *Polygraph*, 7 (3), 173-214.

Summers, W.G. (1939). Science can get the confession. *Fordham Law Review*, 8, 335-354.

United States Army Military Police School. (July 1976). Lesson plan "General question test (GQT) construction.

Weir, R.J., Jr. (1974). In defense of the relevant-irrelevant polygraph test. *Polygraph*, 3 (2), 119-166.

Weir, R.J., Jr. (1976). Some principles of question selection and sequencing for relevant-irrelevant testing. *Polygraph*, 5 (3), 207-222.

Wilkerson, O.W. (1978). The peak of tension tests utilized in the Ashmore kidnapping. *Polygraph*, 7 (1), 16-20.

Wygant, J. (1980). Hypothetical controls. *Polygraph*, 9 (1), 45-48.

Wygant, J. (1979). The relevant-connected control. *Polygraph*, 8 (1), 64-67.

Instrumentation for Presenting a Known Standard Signal to the Electrodermal Activity Channel for Assessing Response Characteristics

Victor L. Cestaro

Abstract

Anecdotal evidence suggested that signals recorded using computerized polygraph instruments with the electrodermal activity (EDA) channel set in the manual mode are substantially different from those recorded with the instrument set in the automatic mode. This had been difficult to confirm since equipment was not available that would generate continuously variable resistance signals of known shape, magnitude, and frequency. Such a device was conceptualized, designed, constructed, tested, and aligned in the laboratory using readily available electronic components. The response characteristics of various computerized polygraph instruments were subsequently analyzed using the device. Results confirmed the verbal reports of differences between signal characteristics in the manual and automatic modes.

Keywords: Electrodermal activity, instrumentation.

Forensic psychophysiologists in the field claimed that the electrodermal activity (EDA) channel on computerized polygraph instruments displayed markedly different responses, depending on whether the signals were recorded in the manual or automatic mode. There was no way to easily verify which signal was the more accurate representation of the physiological changes monitored by the instrument. Most modifications made to the later instruments appeared to be directed at making the equipment easier to use and the output easier to interpret within the context of the psychophysiological detection of deception (PDD).

It was decided that the most rational approach to addressing the linearity issue was to test all instruments, using input signals of known shape, amplitude, and frequency, and then to compare the outputs to the output of an amplifier devoid of any filtering or shaping circuits.

Instrument Design

Concept

The EDA test fixture, hereinafter referred to as the simulator, should be capable of presenting a constantly changing resistance to the input of the polygraph EDA channel. The baseline resistance should be adjustable to simulate a portion of the range of human tonic skin resistance, and dynamically change at a rate equivalent to that of the human phasic response observed during PDD examinations, with an adjustment for center frequency of the phasic response. The signal should also have a known output waveshape in order to provide a subjective assessment of the linearity of the target instruments. It was decided that the device should be capable of presenting sine, triangle, and square waves, with frequencies adjustable from 0.1 to 1 Hz, with an adjustable tonic resistance level of 10K ohms (10,000 ohms) to 100K ohms, and

This project was funded by the Department of Defense Polygraph Institute as DoDPI-P-0013. These results were previously reported in Cestaro (1997). The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. For reprints, write to Dr. Cestaro at DoDPI, Building 3195, Ft. McClellan, Alabama 36205-5114.

a selectable dynamic (corresponding to phasic) response of 5K and 10K ohms. Additionally, provision would be made to accommodate an external input signal to drive the device. The simplest way to control resistance to the input of the EDA channel would be to use a light sensitive resistor (photoresistor) and a controllable light source. However, it was found that simply varying the magnitude of the voltage to a lamp filament, or to a light emitting diode (LED), resulted in a voltage to resistance transfer function that was too nonlinear to be useful. The most linear function, using an LED/photoresistor pair, was obtained during lab tests using a combination of pulse-width and frequency modulation of the LED.

Design

The simulator was designed using a combination of linear and digital integrated circuits (refer to Figure 1). The primary output waveforms are obtained from the 8038 function generator (U1). The 8038 has three simultaneous outputs; sine, square, and triangle waveforms. Only the square wave output is nonlinear, having a voltage swing from ground to Vdd (the input power supply voltage level). The other two outputs are linear with maximum amplitudes approximately equal to Vdd/4 for the sine wave and Vdd/3 for the triangle. The frequency of all three waveforms is simultaneously controlled by R1, which is adjustable from 0.1 to 1.0 Hz. Waveform symmetry is controlled by R2, and is adjusted for best sine or triangle symmetry at 0.5 Hz. Potentiometers R4 and R5 are used to adjust the sine output for minimum distortion (third harmonic and peaks). All three outputs are alternating current (AC) coupled to a 4066 analog switch (U2) through level adjustment potentiometers (R6, R7, and R8). Selection of signals switched through the 4066 is accomplished by a 4017 decade counter (U7) used as a stepping switch, clocked manually by an LM555 timer (U5) wired as a monostable multivibrator (one-shot) triggered by a front-panel momentary push-button switch (S2) connected to a differentiating network in front of the one-shot.

The four outputs of the 4066 are fed through 10K ohm resistors to a summing junction at the input of an LM324 operational amplifier (U3A), with adjustable gain and output offset, which in turn is direct current (DC) coupled to the control input of U8, an LM555 configured as a voltage controlled oscillator (VCO). The offset control (R9) is used to set the starting frequency of the VCO by adjusting the trough of the triangle wave to 1.0 volts (DC level). The range of the VCO control voltage is set for 1.0 to 4.0 volts, with the 4.0 volt upper limit (triangle wave peak) determined by the gain setting of the LM324 amplifier (U3A). The three waveforms are adjusted for equal amplitudes at the output of U3A by R6, R7, and R8. The linearity of the voltage to frequency transfer of the VCO was found to be satisfactory for this application. The VCO output is connected to the base of a 2N2222A NPN transistor which functions as a pulse modulated current sink for the LED wired to the collector of the transistor. Source current to the LED is determined by the range switch (S5), which can select one of two current ranges (through R12 and R13) and hence, the resistance change seen at the photoresistor. The voltage for U8 and the modulator transistor is held constant by an LM7805 voltage regulator (VR1). The response time of the photoresistor is long enough to cause it to act as a low-pass/high-reject filter to smooth out the 25-50 Khz (25,000 to 50,000 Hz) light pulses from the LED. The change in resistance, which follows the low frequency control voltage waveform at pin 5 of U8, is ripple-free. Potentiometer R16 is used to adjust the baseline resistance level (tonic EDA).

An additional LM555 (U6) is wired as a one-shot to apply a reset pulse to U7 so that when power is applied the device starts in the OFF state (no signal to the VCO). A 14528 retriggerable one-shot (U10) is used to detect the presence of pulses at Q1 collector and lights a diagnostic lamp to enunciate the loss of pulse modulation to the LED. U9C and U9D are used to detect loss of regulation of the +5 volt supply to the VCO, and will illuminate a red front panel lamp if the voltage falls to less than 4.7V or rises above 5.3V. Within the normal voltage range, a green lamp is illuminated. An 8871 open-collector driver

(U4) is used to illuminate the various front panel function and diagnostic LEDs. An external function generator or other modulating device may be used to drive the system through the external input at U3B. Two 1N4148 diodes in series (D1 and D2) are used to clamp the input at approximately one volt so that the baseline of the external signal is roughly equivalent to the three internal signal baselines in order to avoid large excursions on the display of the unit under test when the signal source is switched.

The simulator was assembled on a general-purpose component PC board purchased from Radio Shack (#276-147A), and installed in a plastic enclosure also purchased from Radio Shack. The general component layout is shown in Figure 2.

Initial alignment of the simulator was accomplished using an oscilloscope to monitor the output of the 8038 with switch S4 in the test position (higher frequency). The final alignment in the 0.1 to 1 Hz operating frequency range was done using a data acquisition board and software installed in an IBM compatible 486 personal computer. In order to convert the resistance output of the simulator to voltage for the data acquisition system, a resistance to voltage converter was built (Figure 3).

Method

Apparatus

A four-channel oscilloscope (Model 2247A, Tektronix, Beaverton, OR) was used to test and align the simulator prior to testing the polygraph instruments, and for monitoring the simulator output during testing. An IBM compatible 486 computer with an internal Dataq analog to digital converter (ADC), using CODAS data acquisition and display software (Dataq Instruments Inc., Akron, OH) was used to record the output of the simulator to establish a measurement standard. Three computerized polygraph instruments were tested for response linearity: (a) Lafayette LX2000, (b) Stoelting CPS (software version 2.14D), and (c) Axciton (Interface Box Version S7.1, software release 4.9). Input to the CODAS adc was not filtered. A flatbed scanner (Model IIcx, Hewlett-Packard, Palo Alto, CA), attached to an IBM compatible 486 computer, was used to scan polygraph charts and save them on magnetic media as computer image files.

Procedure

The EDA channel leads on each instrument were individually connected to the EDA output of the simulator. The simulator frequency control was adjusted for an output frequency of 0.1 Hz. The simulator tonic EDA level control was adjusted to 50K ohms. The simulator range switch was set to 5K or 10K, depending on the instrument being tested, to get a EDA trace that would remain within the limits of the EDA channel amplifier linear region. Three one minute data epochs were collected for sine, square, and triangle wave outputs from the simulator for each instrument. Data were subsequently printed on paper charts for each of the three instruments. Finally, the paper charts were scanned into tagged image format (TIF) computer files for post analysis using the Hewlett-Packard scanner.

Results

The output of the EDA channels on the Axciton and the Lafayette instruments differed between the AUTO and MANUAL modes of operation. The Stoelting system has no mode switch. Figures 4, 6, and 8 show the output of the Axciton instrument in the AUTO mode. Figures 5, 7, and 9 depict the same instrument in the MANUAL mode. The positive baseline shift apparent in Figures 5, 7, and 9 was not seen on the input signal, but is related to some function within the Axciton polygraph instrument. This baseline shift was not observed on the other two instruments, nor on the CODAS system used to calibrate the simulator. Additionally, nonlinear transduction of the triangle and square wave resistance signals was observed on the Lafayette LX2000 and the Axciton systems in the AUTO mode (Figures 6, 8, 12, and 14). Signal linearity was good for the sine wave on the Stoelting CPS, with some distortion on the triangle and square waves (Figure 16). Linearity was good for the sine wave in the AUTO and MANUAL modes on the LX2000

and the Axciton (Figures 4, 5, 10 and 11). Some distortion was observed on the triangle and square waves in the MANUAL mode on the Axciton and Lafayette (Figures 7, 9, 13, and 15). In all cases, the instrument sensitivity (gain) was not adjusted when changing from AUTO to MANUAL. Amplitude differences observable in the figures are due to internal differences between the two modes of operation.

Discussion

The Lafayette LX2000 and the Axciton instruments displayed nearly identical changes in their output waveforms when switched between the AUTO and MANUAL modes of operation. However, the Axciton instrument demonstrated a pronounced positive baseline (tonic) shift in the MANUAL mode. The manufacturer is apparently aware of this situation, and customers have been advised to use the instrument only in the AUTO mode. The shift was not seen on the other two instruments. The Stoelting instrument output resembled the other two instruments in the MANUAL mode. The minor signal distortion on the triangle and square waves observed in the MANUAL mode can be attributed to the effects of rejection of high frequencies by the built-in low pass filters in the instruments.

When the AUTO mode was selected on the LX2000 and the Axciton, the output signal deviation from the input signal was most apparent when the input was either a triangle or square wave, and appeared to be differentiated. This may be due to AC coupling on the input to reduce the amplifier response to DC baseline, or tonic skin resistance changes. The problem was less apparent when the input signal was a sine wave. This characteristic distortion could conceivably be considered non-problematic in the field because the EDA rise and fall times are more closely represented by the rise and fall times of the sine wave rather than by the times associated with the triangle and square waves. However, the *degree* of the observable differences among the fast rise time signal responses, with emphasis on the triangle, indicates a disparity among instruments which is related to gross differences in the internal filtering of the physiological signal being measured (see Figures 6, 12, and 16). Filter characteristics were found to differ considerably among manufacturers, and also between different models from the same manufacturer. Additionally, amplifier gain parameters and analog to digital conversion (ADC) rates vary considerably among manufacturer's products.

Finally, there appear to be no established standards for physiological measurements within the polygraph instrument manufacturing industry, as is evident from the instrument specifications supplied by the manufacturers (Table 1). The filter characteristics for the LX2000 and LX3000 are markedly different, as are the differences among instruments from the three manufacturers. It may be prudent for the manufacturers to collectively select and agree upon channel specifications for each component and provide amplifier outputs which are accurate representations of the physiology being measured. Basic standardization may become increasingly more important as computerization is used to take advantage of scoring algorithms for automated decision-making. Currently, scoring algorithms are not portable among the different instruments, nor are the basic algorithms available for analysis by physiologists and computer scientists working in the PDD discipline. Until there is some standardization, there is no way that the most efficacious algorithms can be easily implemented in the native code for each instrument so that there will be confidence in inter-instrument scoring reliability.

References

Cestaro, V. L. (1997). Instrumentation for presenting a known standard signal to the EDA channel for assessing response characteristics of selected polygraph instruments. (Report No. DoDPI96-R-0004). Fort McClellan, AL: Department of Defense Polygraph Institute.

Figure Captions

Figure 1. Schematic diagram of the EDA test device (simulator).

Figure 2. Component layout of the EDA simulator.

Figure 3. Resistance to voltage converter for the CODAS system.

Figure 4. Axciton chart showing a 0.1 Hz sine wave on the EDA channel in AUTO mode.

Figure 5. Axciton chart showing a 0.1 Hz sine wave on the EDA channel in MANUAL mode.

Figure 6. Axciton chart showing a 0.1 Hz triangle wave on the EDA channel in AUTO mode.

Figure 7. Axciton chart showing a 0.1 Hz triangle wave on the EDA channel in MANUAL mode.

Figure 8. Axciton chart showing a 0.1 Hz square wave on the EDA channel in AUTO mode.

Figure 9. Axciton chart showing a 0.1 Hz square wave on the EDA channel in MANUAL mode.

Figure 10. LX2000 chart showing a 0.1 Hz sine wave on the EDA channel in AUTO mode.

Figure 11. LX2000 chart showing a 0.1 Hz sine wave on the EDA channel in MANUAL mode.

Figure 12. LX2000 chart showing a 0.1 Hz triangle wave on the EDA channel in AUTO mode.

Figure 13. LX2000 chart showing a 0.1 Hz triangle wave on the EDA channel in MANUAL mode.

Figure 14. LX2000 chart showing a 0.1 Hz square wave on the EDA channel in AUTO mode.

Figure 15. LX2000 chart showing a 0.1 Hz square wave on the EDA channel in MANUAL mode.

Figure 16. Stoelting CPS charts depicting 0.1 Hz sine, triangle, and square waves on the EDA channel output.

Table 1
Amplifier Characteristics of Various Polygraph Instruments

Mfr.	Hi-pass filter (-3 dB)	Lo-pass filter (-3 dB)	Filter rolloff dB/octave	Electrode current uAmps*	Electrode voltage	Amplifier gain	Amplifier slew rate	AD sample rate
				EDA channel				
LX3000	0 Hz	1 Hz	6	5.784	2.975 V	1 - 11	28V/mSec	120 Hz
LX2000	0 Hz	97 Hz	12	9.018	4.55 V	5 - 44	.01V/uSec	1000 Hz
Axciton	.08 Hz	5 Hz	20	2	1.0 V	5 - 40	---	30 Hz
Stoelting	---	6 Hz	18	---	---	1639	---	48 Hz
				Cardio channel				
LX3000	0 Hz	48 Hz	6			356	28V/mSec	120 Hz
LX2000	0 Hz	111 Hz	12			802	.05V/uSec	1000 Hz
Axciton	0 Hz	60 Hz	20			10 - 600	---	120 Hz
Stoelting	---	30 Hz	10			174	---	48 Hz
				Respiratory (pneumo) channel				
LX3000	0 Hz	16 Hz	6			196	28V/mSec	120 Hz
LX2000	0 Hz	111 Hz	12			402	.05V/uSec	1000 Hz
Axciton	0 Hz	15 Hz	20			10 - 500	---	30 Hz
Stoelting	---	15 Hz	10			60	---	48 Hz

Note. *Lafayette and Axciton assume 500K nominal skin resistance in their EDA electrode current data.

Figure 1
Schematic diagram of the EDA test device (simulator).

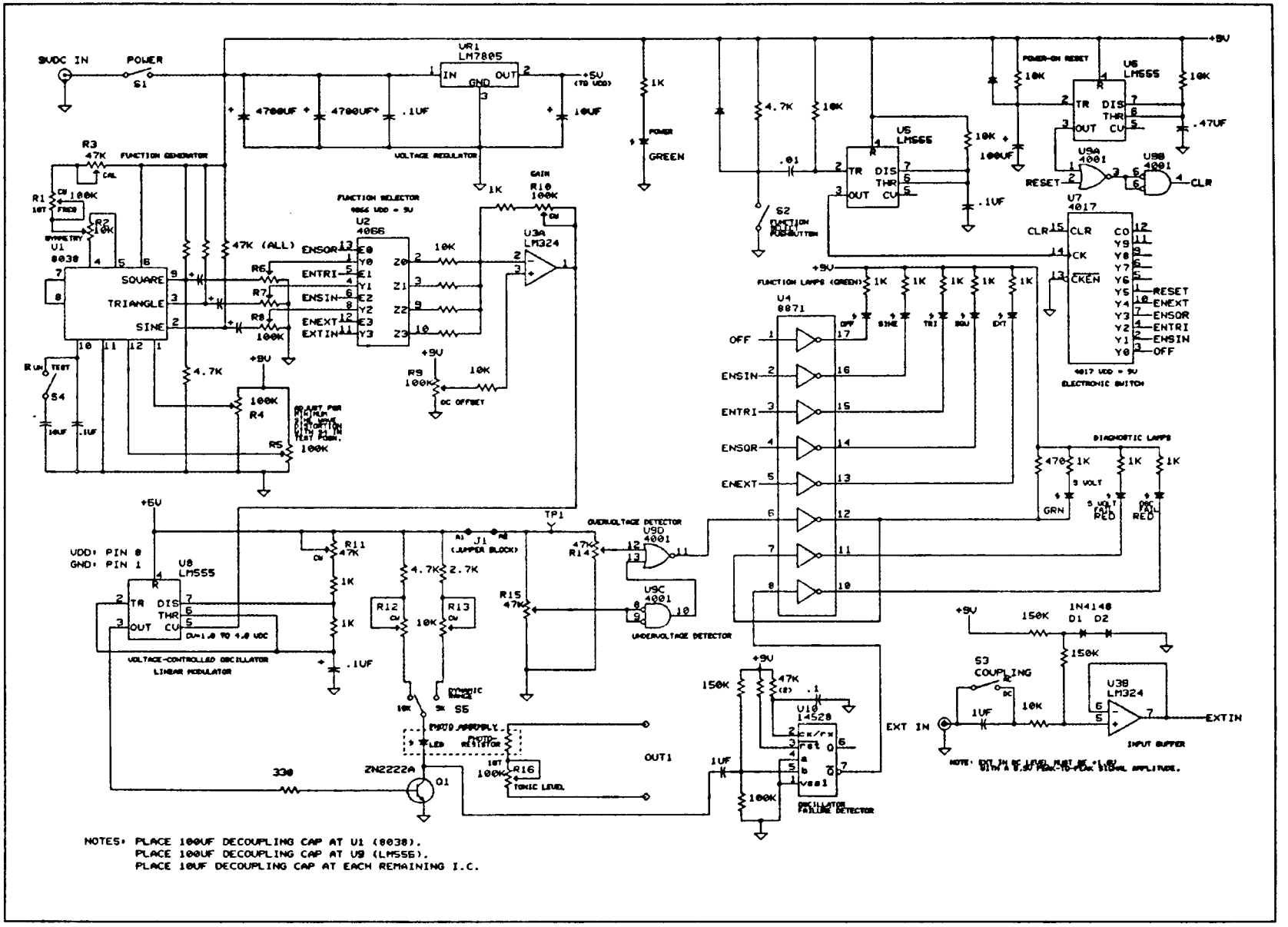


Figure 2
Component layout of the EDA simulator.

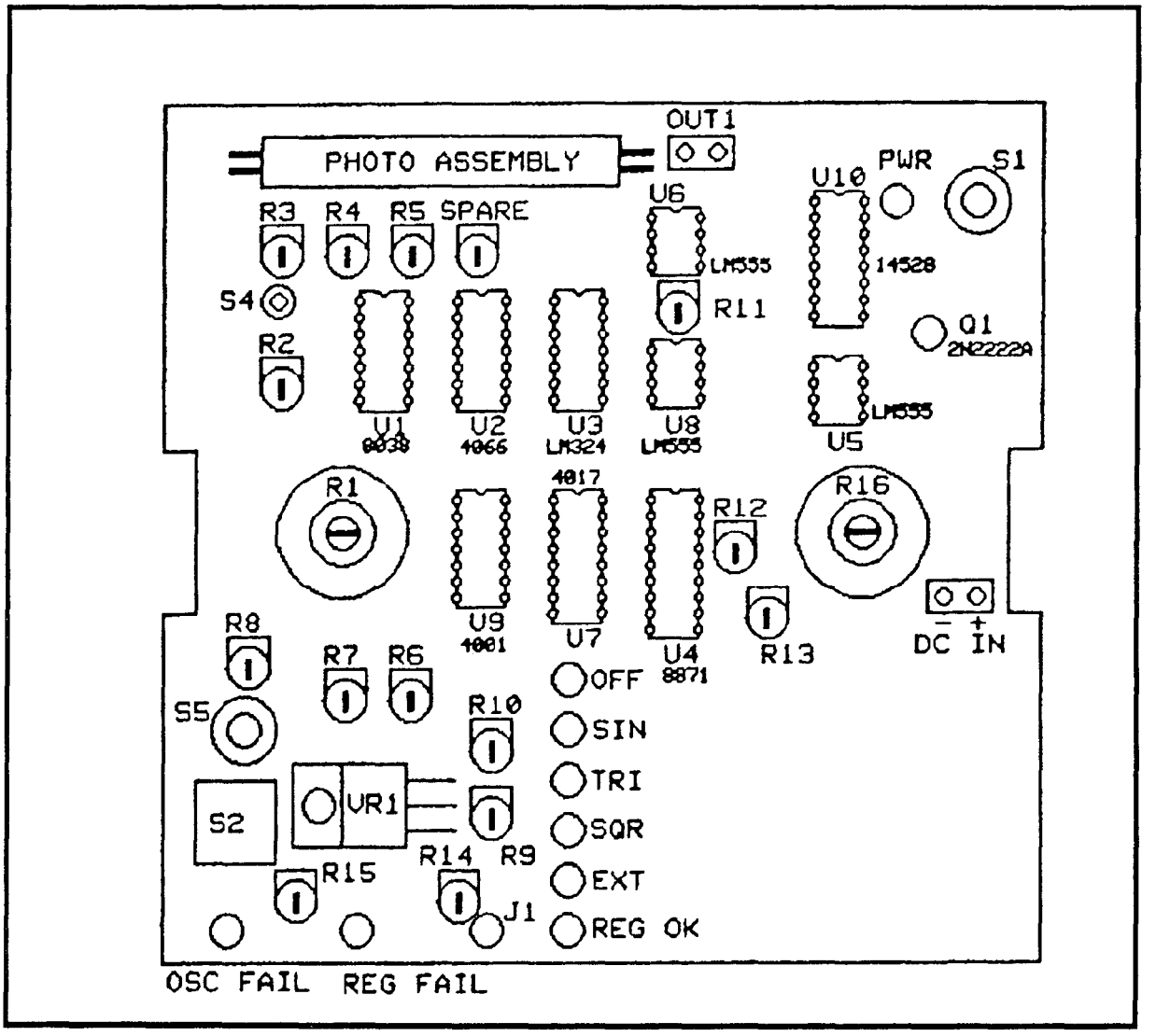


Figure 3
Resistance to voltage converter for the CODAS system.

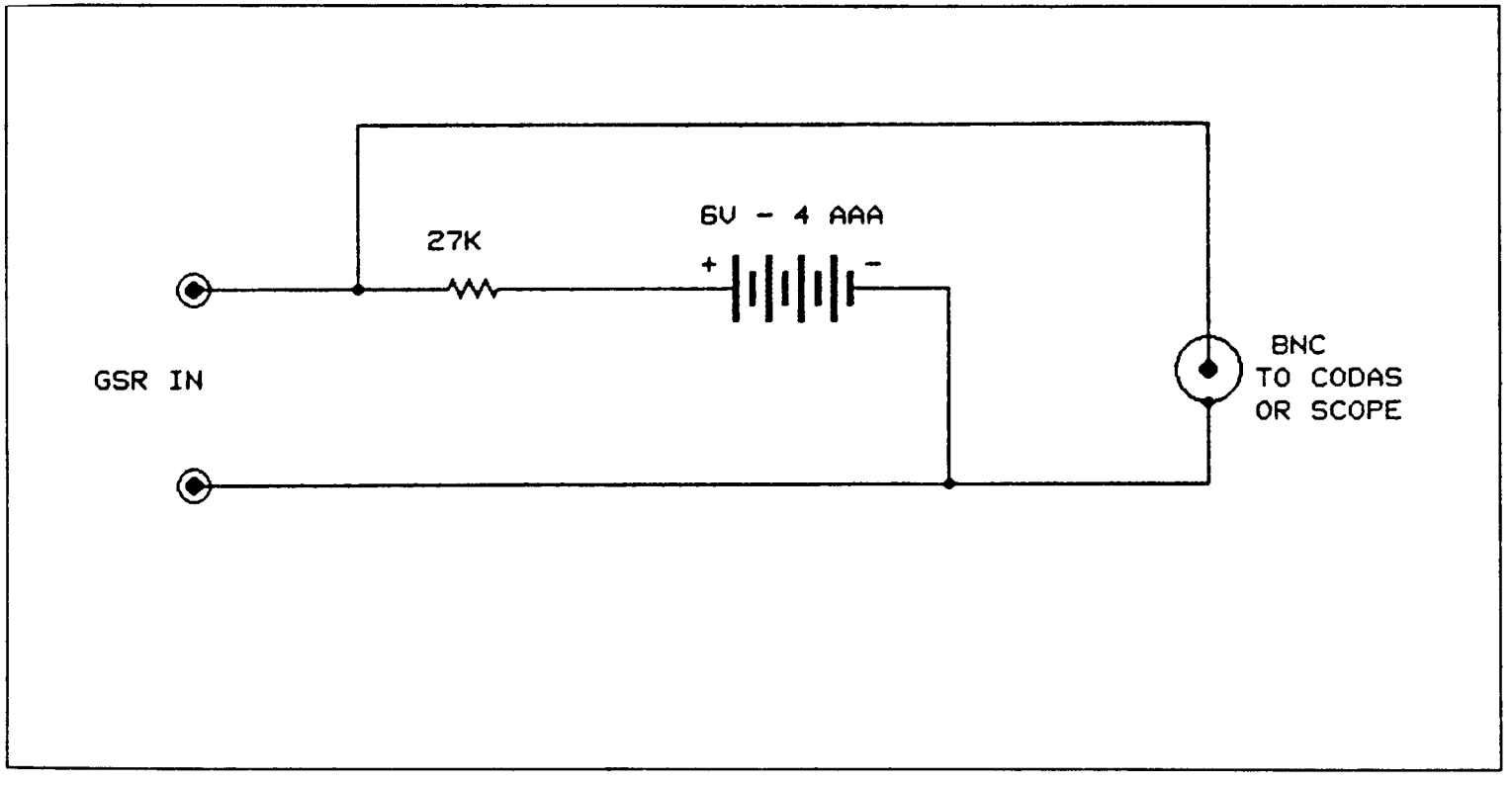


Figure 4
Axciton chart showing a 0.1 Hz sine wave on the EDA channel in AUTO mode.

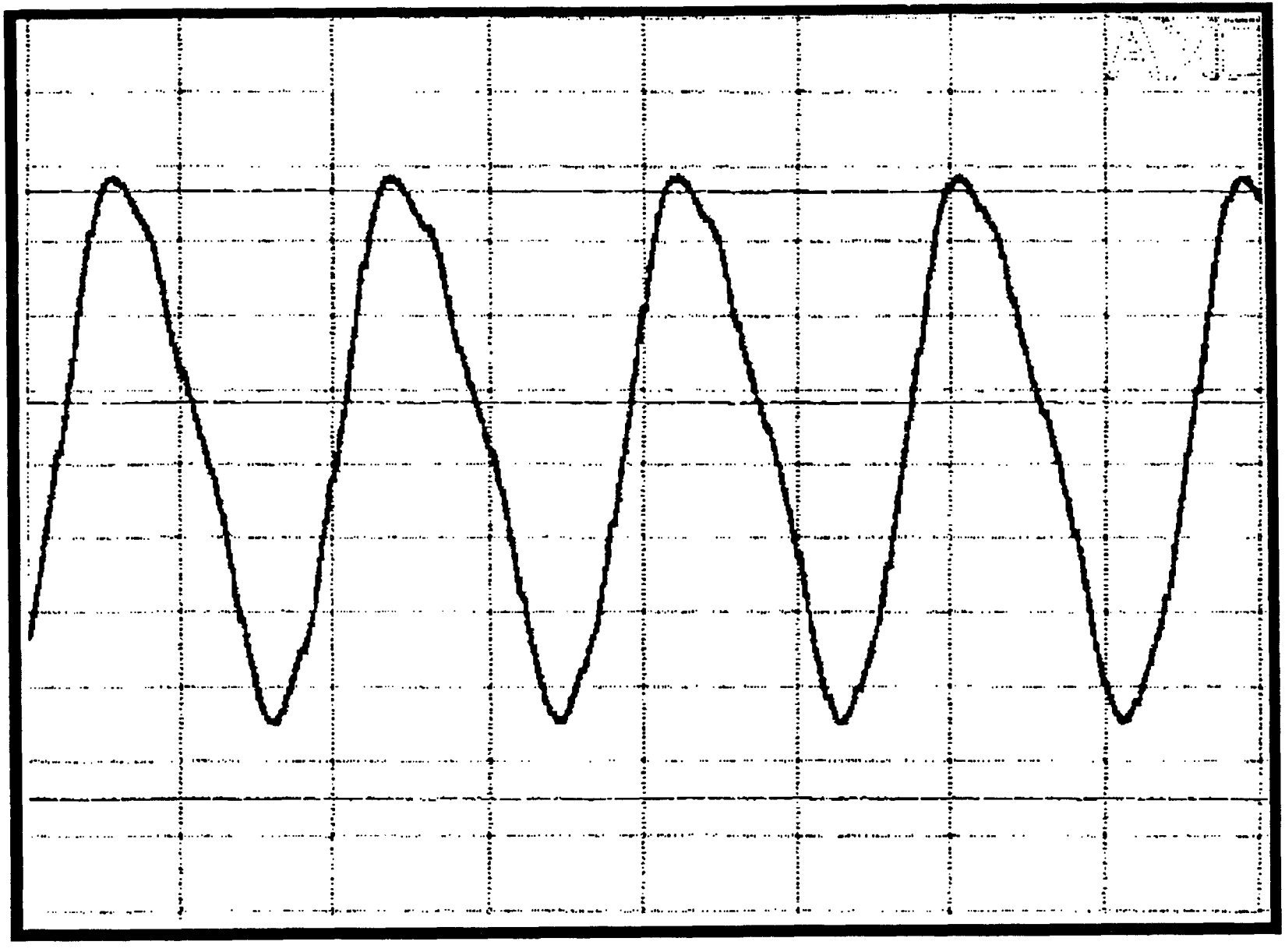


Figure 5
Axciton chart showing a 0.1 Hz sine wave on the EDA channel in MANUAL mode.

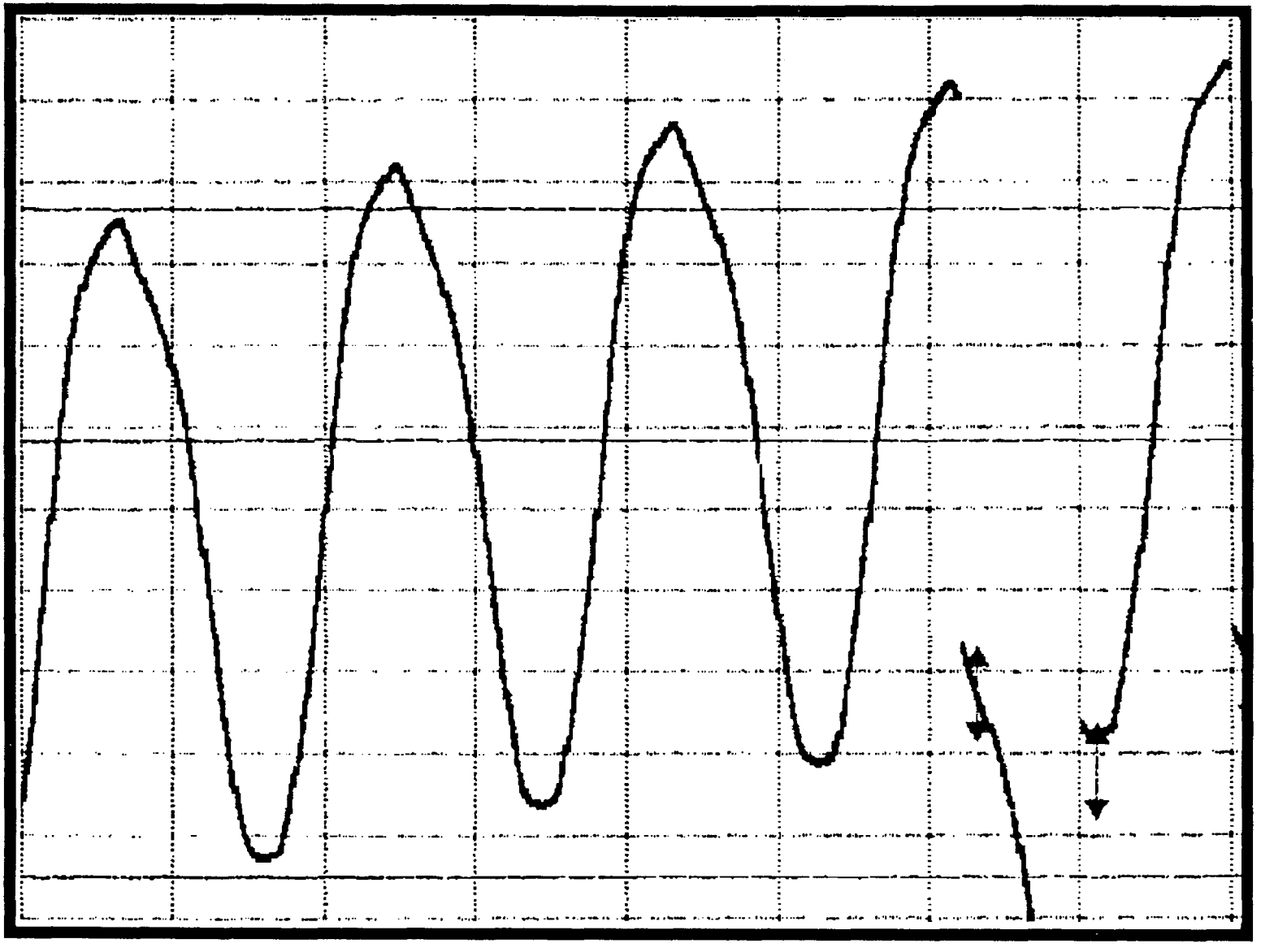


Figure 6
Axciton chart showing a 0.1 Hz triangle wave on the EDA channel in AUTO mode.

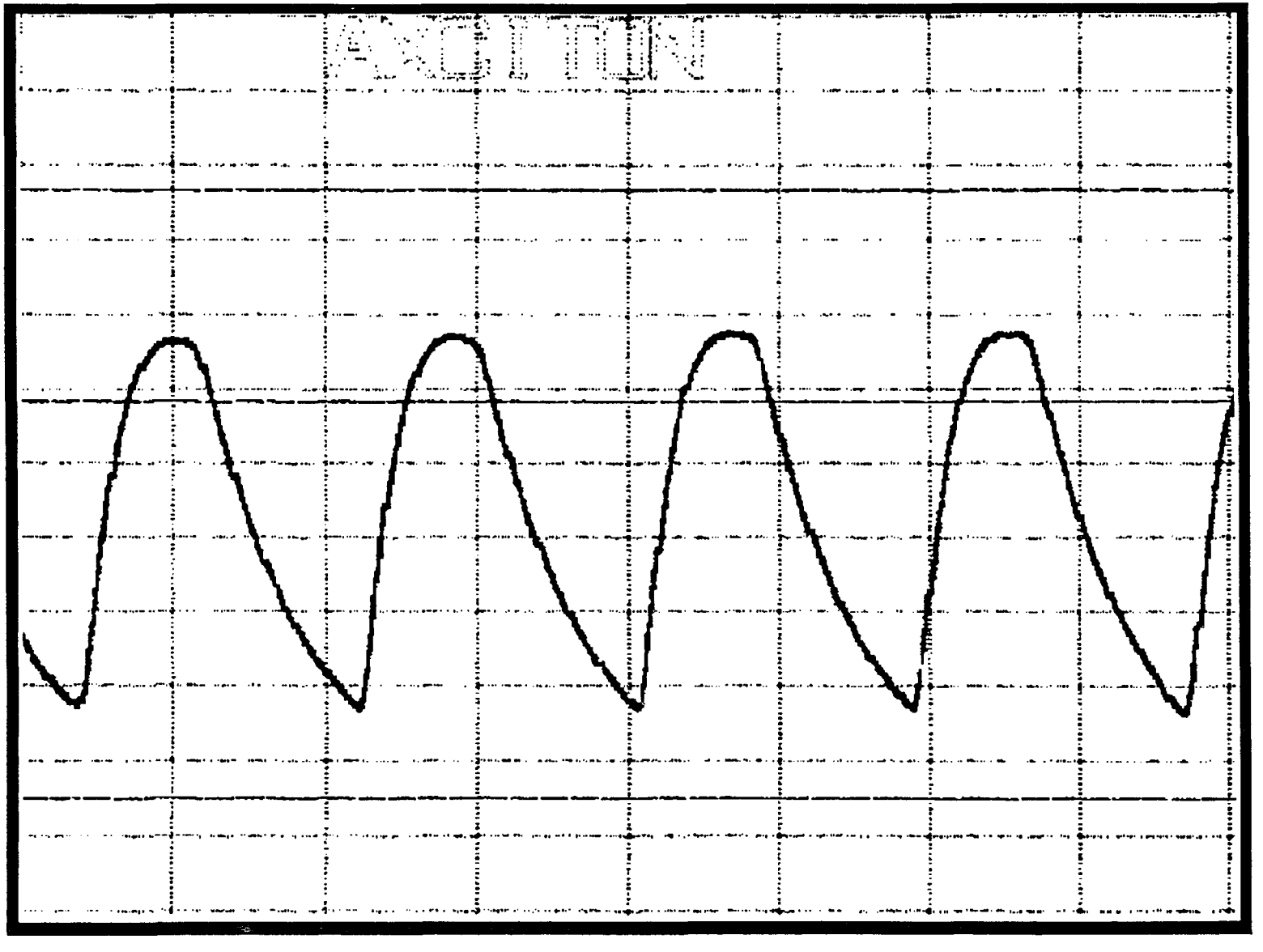


Figure 7
Axciton chart showing a 0.1 Hz triangle wave on the EDA channel in MANUAL mode.

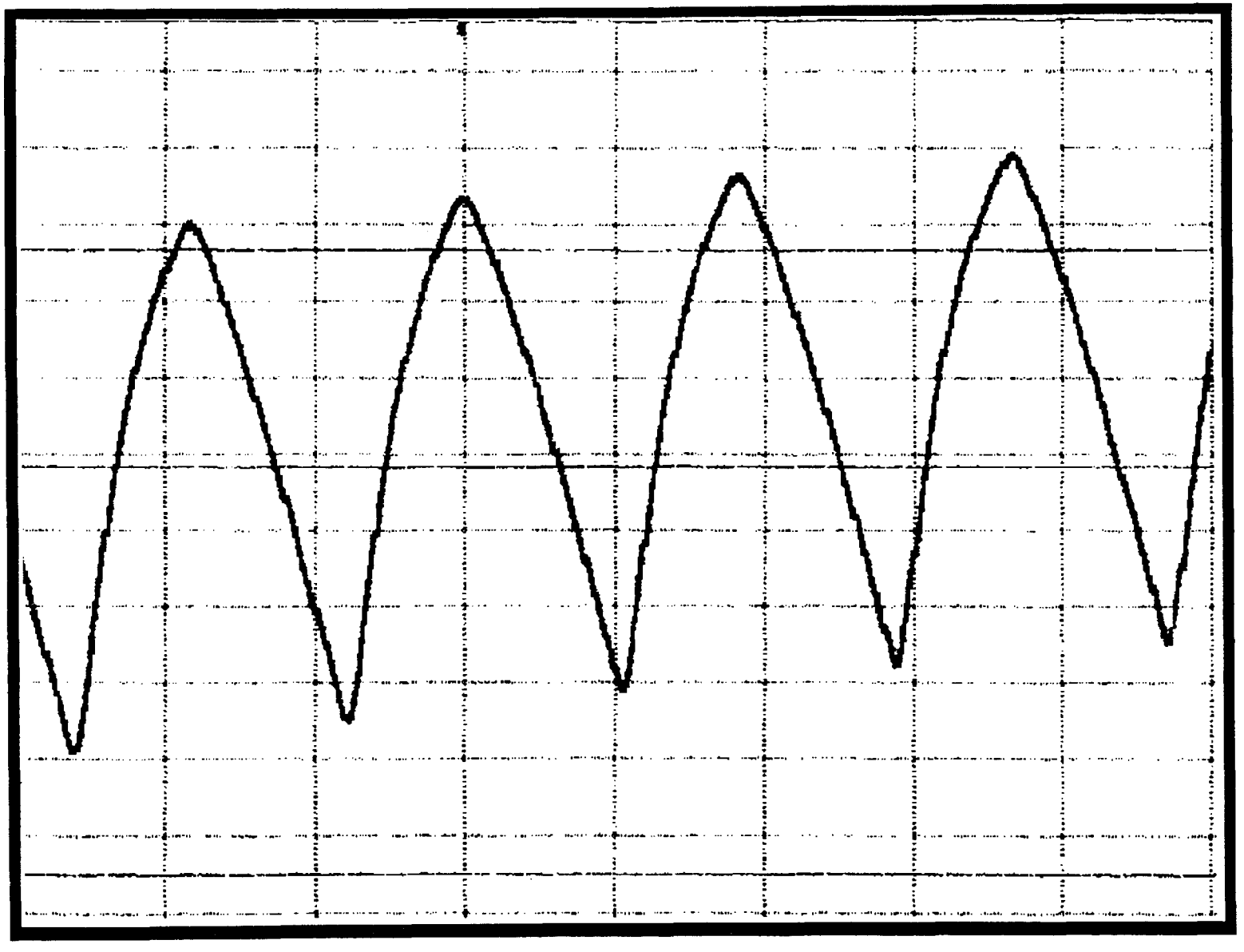


Figure 8
Axciton chart showing a 0.1 Hz square wave on the EDA channel in AUTO mode.

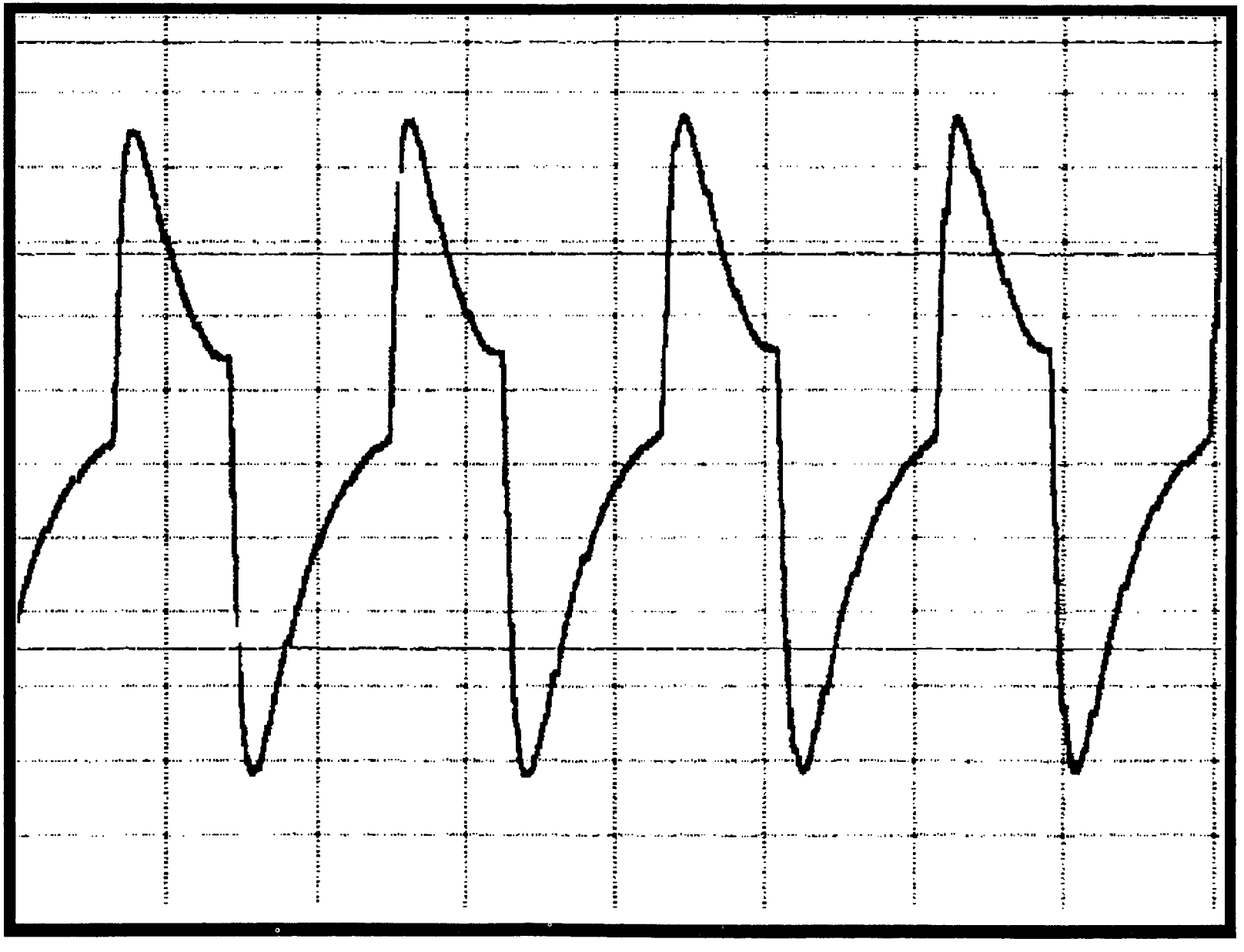


Figure 9
Axciton chart showing a 0.1 Hz square wave on the EDA channel in MANUAL mode.

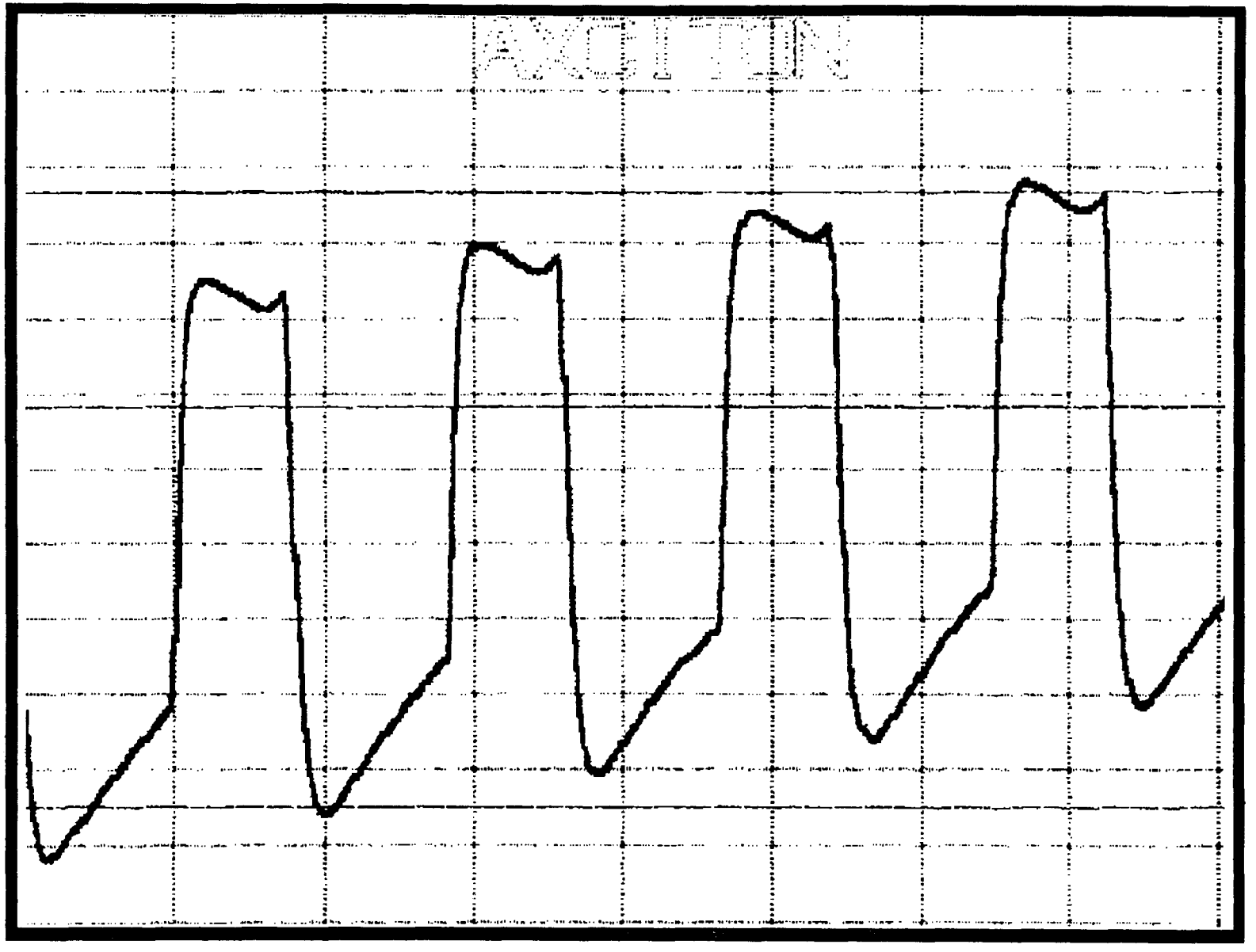


Figure 10
LX2000 chart showing a 0.1 Hz sine wave on the EDA channel in AUTO mode.

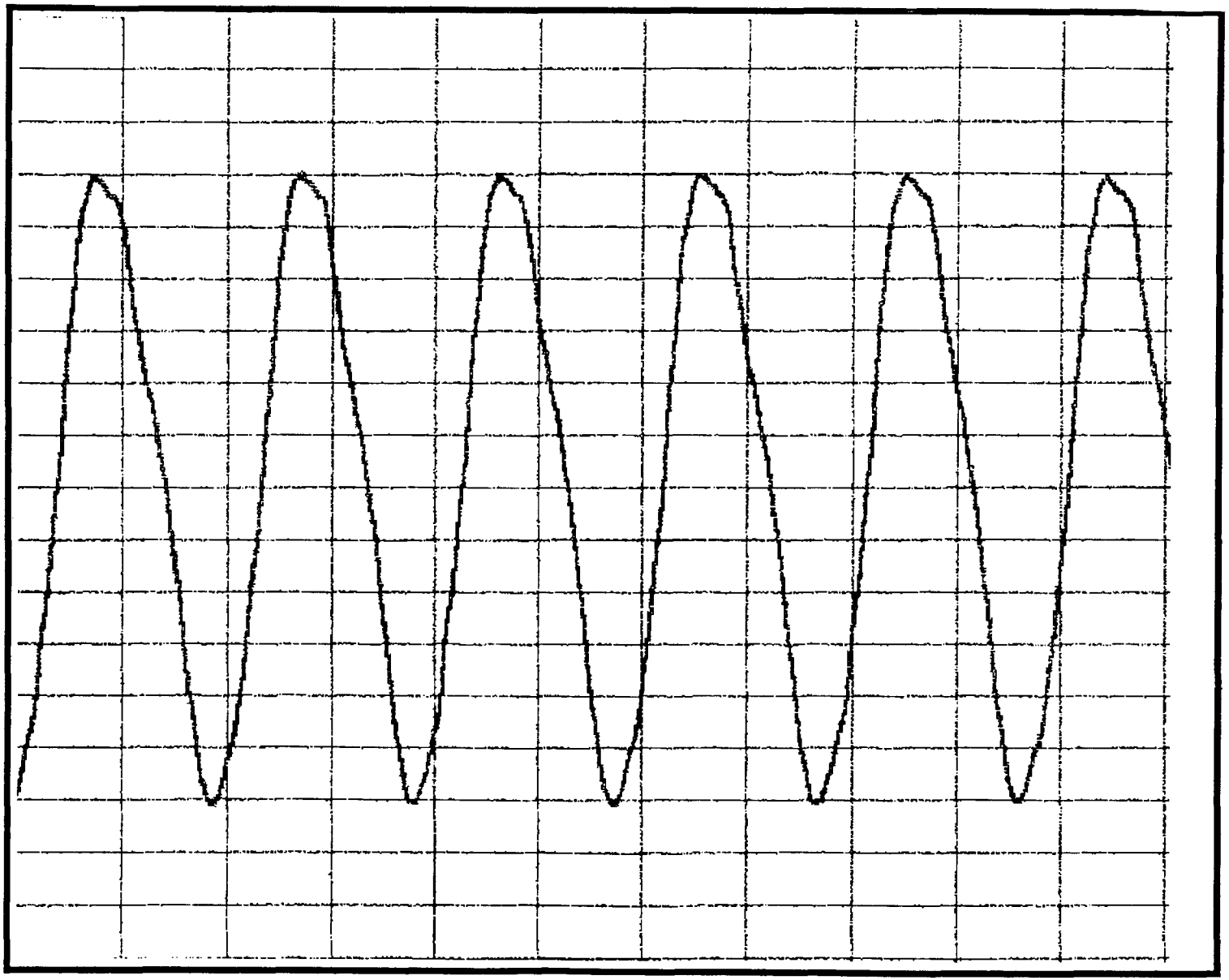


Figure 11
LX2000 chart showing a 0.1 Hz sine wave on the EDA channel in MANUAL mode.

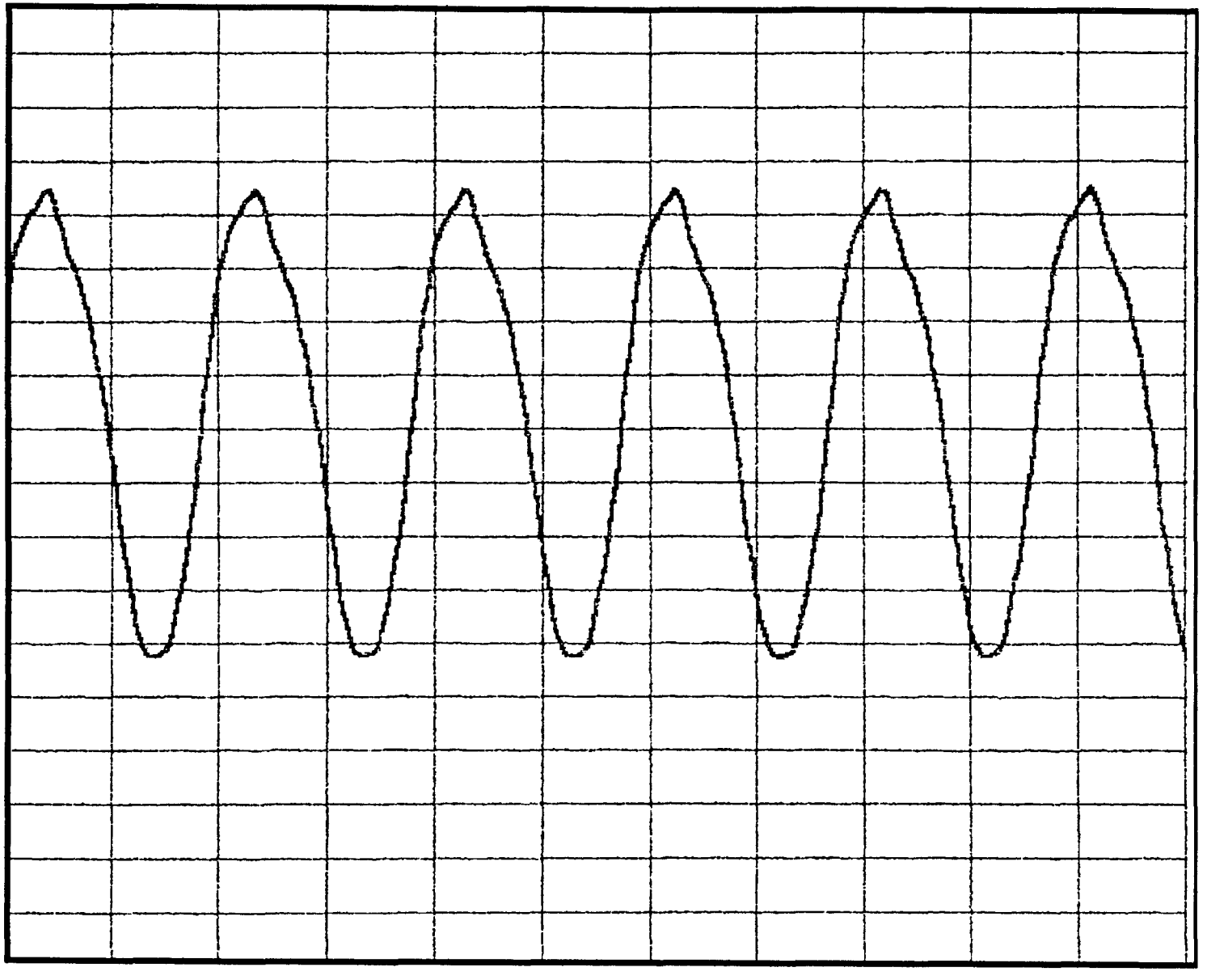


Figure 12
LX2000 chart showing a 0.1 Hz triangle wave on the EDA channel in AUTO mode.

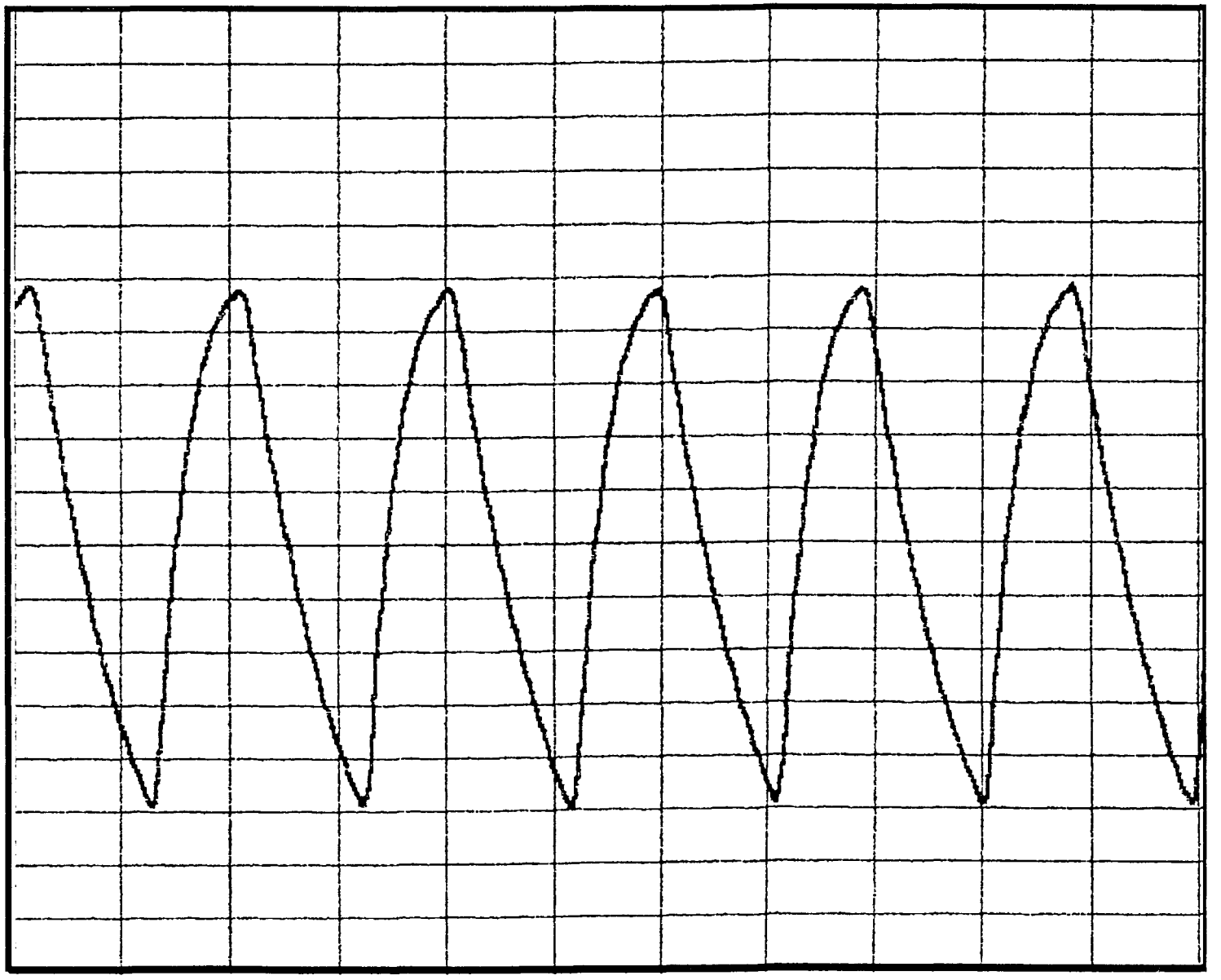
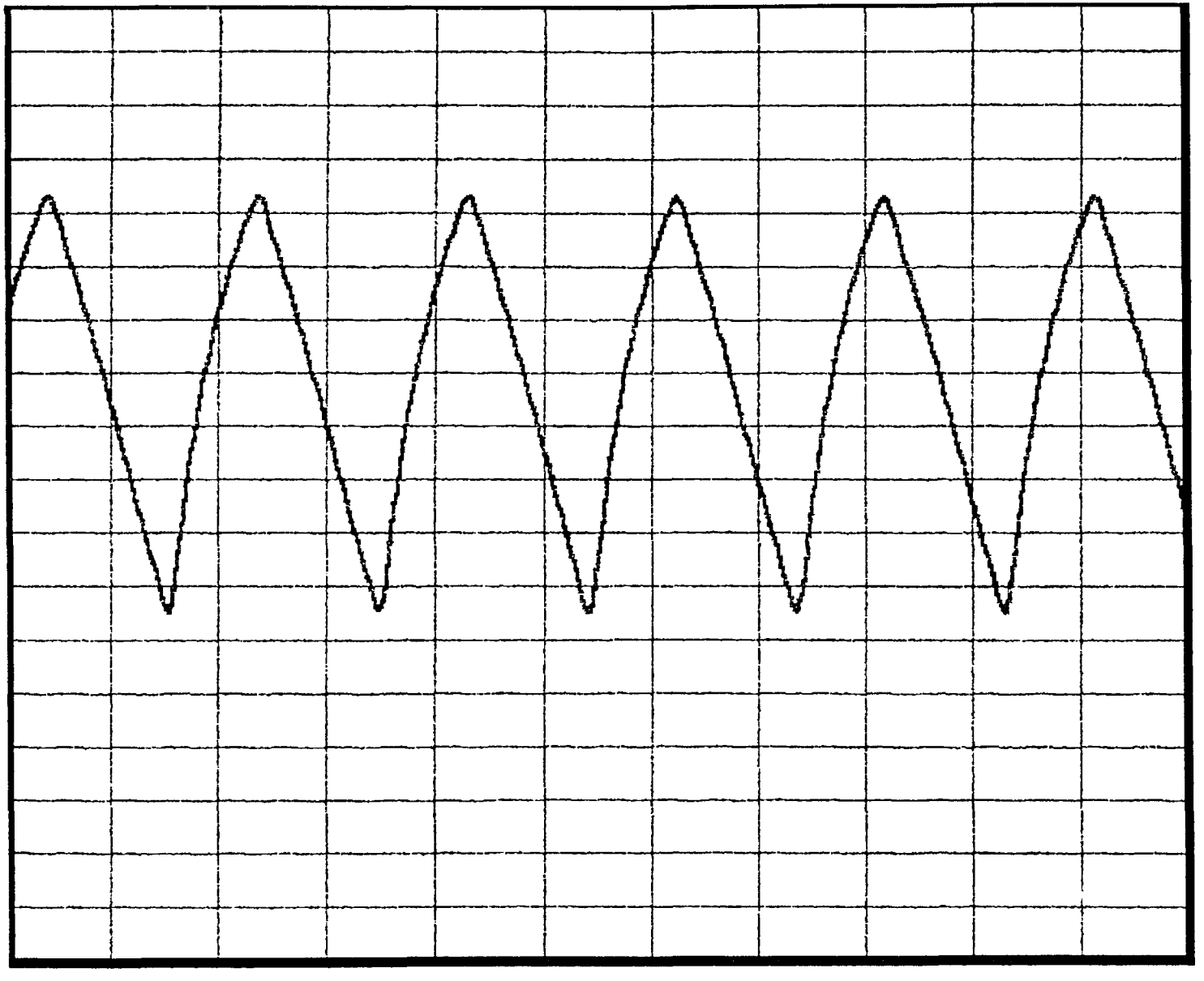


Figure 13
LX2000 chart showing a 0.1 Hz triangle wave on the EDA channel in MANUAL mode.



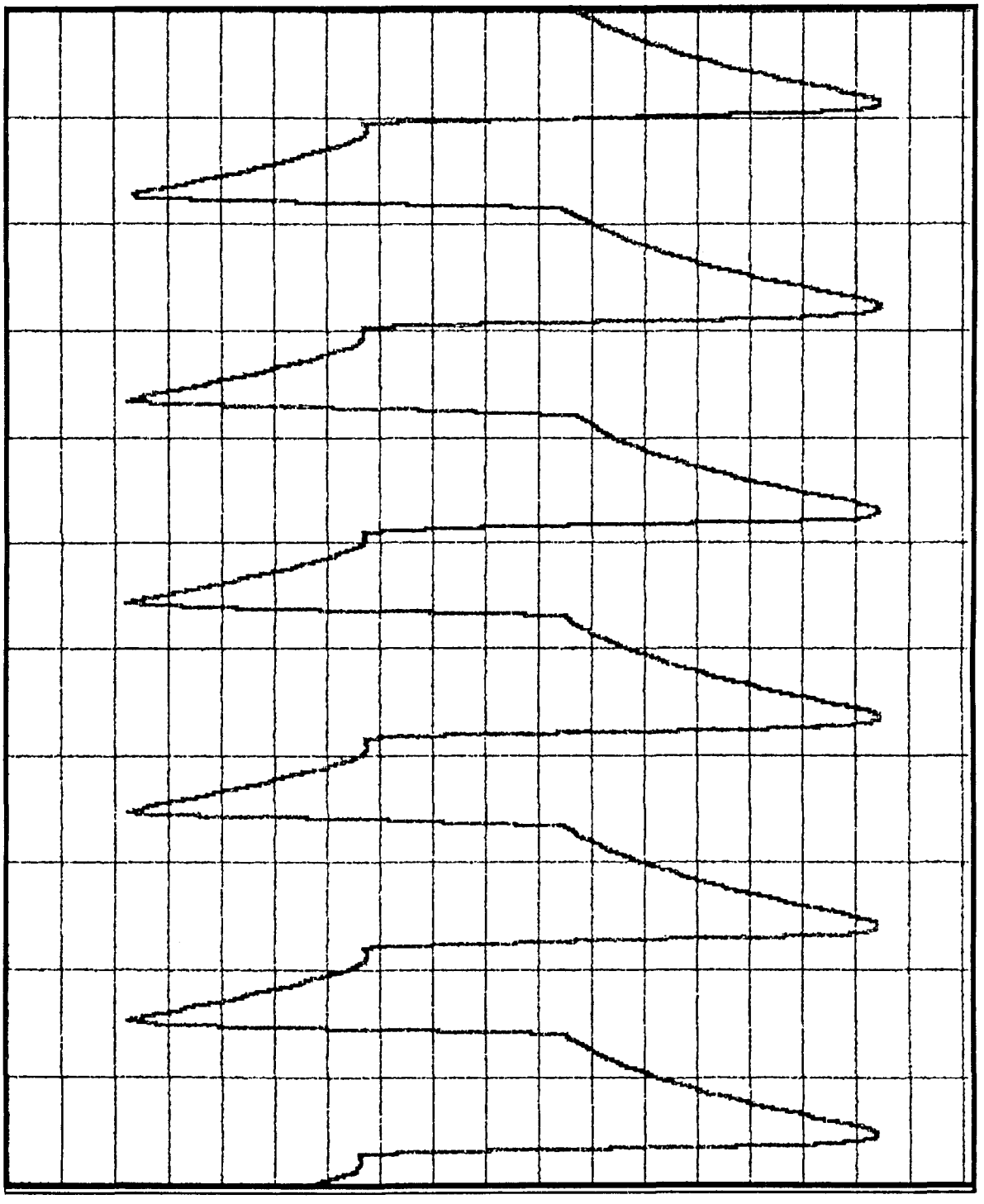


Figure 14
LX2000 chart showing a 0.1 Hz square wave on the EDA channel in AUTO mode.

Figure 15
LX2000 chart showing a 0.1 Hz square wave on the EDA channel in MANUAL mode.

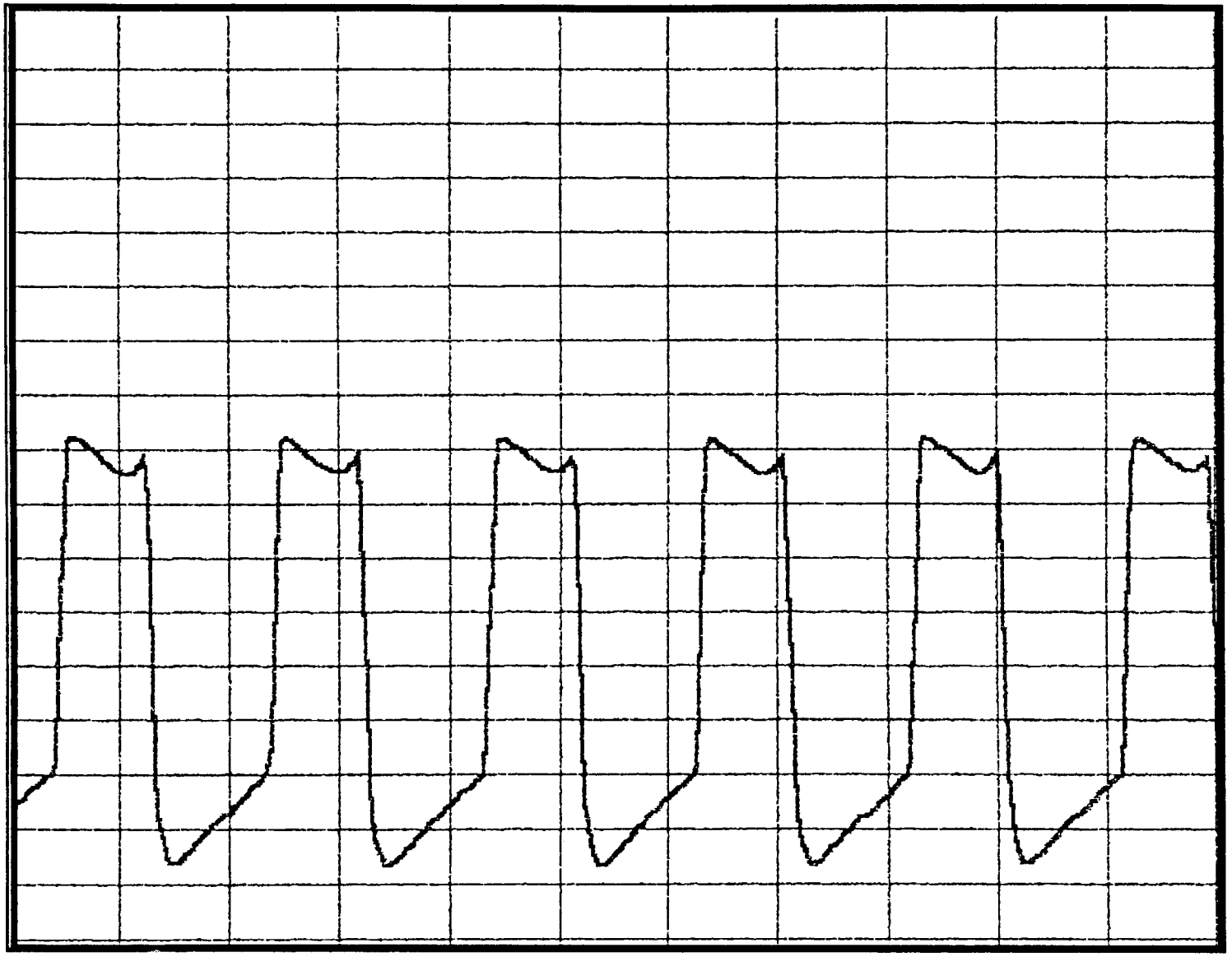
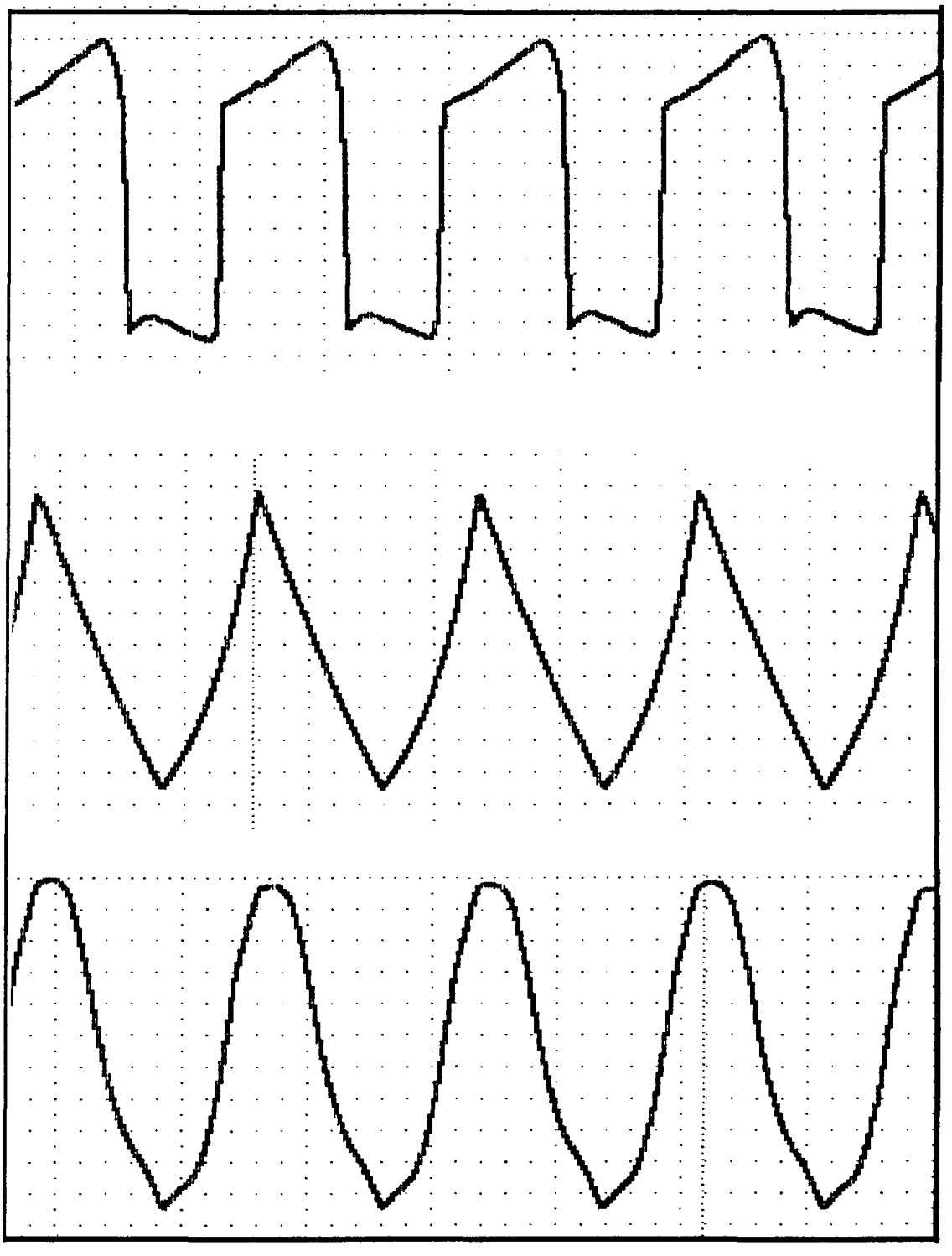


Figure 16
Stoelting CPS charts depicting 0.1 Hz sine, triangle, and square waves on the EDA channel output.



A Comparison of 3- and 7-position Scoring Scales with Laboratory Data

Donald J. Krapohl

Abstract

The 7-position scale is considered the mainstay of numerical analysis in the field of psychophysiological detection of deception (PDD). A similar method, the 3-position scale, is also widely practiced. Neither method has been subjected to a thorough assessment, and the 3-position method has hardly been investigated at all. Moreover, to a lesser degree than the 7-position scale, the appropriate cutting scores of the 3-position scale have not been explored. In the present effort we systematically evaluated the efficacy of the 3-position scale at different decision thresholds. It was determined that a cutting score of +/-4 for the 3-position scoring system had the least variation from, and was statistically equivalent to, the widely accepted +/-6 cutting score of the 7-position scale when applied to single-issue test formats. It was also noted that the highly experienced raters in this study rarely used the full range of available values in the 7-position scale, employing the narrower range of the 3-position scale for about 90% of the question comparisons. In addition, a post hoc analysis of the 7-position scores found that, consistent with other research, the spot score rule increased true positives at a cost of higher false positives. The problem of identifying optimum cutting scores was also addressed.

Keywords: 7-position scoring, 3-postion scoring, scoring, spot scores

In the discipline of psycho-physiological detection of deception (PDD) there have been two principle methods of arriving at decisions of deception or no deception in the field. In one method, polygraph examiners render decisions based on subjective evaluations of responses to relevant and comparison questions (sometimes called "control questions"), and include information such as case background, verbal indicators, and behavioral analysis in their assessment model. The second method, the numerical approach, entails the assignment of numerical values to physiologic data. Examiners using numerical evaluation strive to exclude extrapolygraphic information from their decision rules, and the numerical method has found current favor in the polygraph, scientific and legal communities.

Numerical scoring systems for PDD data can be traced back to at least the 1930s (Winter, 1936), and several systems have since emerged (Lee, 1943, 1953; Hathaway & Hanscom, 1958, Lykken, 1958; Gordon & Cochetti, 1987; Honts & Driscoll, 1988). These systems categorize examinees as deceptive or truthful based on within-subject comparisons of physiological responses scored with a semi-objective method. The most prominent PDD scoring system in current usage is called the 7-position system, first described by Cleve Backster (1962). While there are some variants, the orthodox

The author is a polygraph research program manager in federal service, and a member of the American Polygraph Association. The conclusions expressed in this paper do not necessarily represent the views of the US Government or the American Polygraph Association.

Acknowledgments: The author is grateful to Dr. Andrew Dollins of the Department of Defense Polygraph Institute for providing the scoring data used in this study,. and for editing the manuscript. Thanks also go to Dr. John Moore and Dr. Victor Cestaro for statistical guidance. The editorial reviews of Dr. Frank Horvath and Dr. William Yankee were also very helpful and appreciated.

7-position scoring system entails the assignment of whole number values between -3 and +3 to differential response patterns. The further from zero the score, the greater the differences in the sizes of the responses to relevant and comparison questions. By convention, larger responses occurring to relevant questions are assigned negative scores, while larger responses occurring to the comparison questions are assigned positive values. Equal responses are assigned zeros. Each physiological parameter for every relevant/comparison question pairing receives one of these values on each test. For example, if there were two relevant questions presented on three tests using a polygraph with the standard three parameters, there would be 2X3X3, or 18 comparisons, each requiring the assignment of a value between -3 and +3. At the end of all testing the values are tallied. Decision rules are dictated by the test format, and thresholds are different for each type of format. Some scoring methods have thresholds for each relevant question, while other scoring systems compare a cutting score with the single number produced by summing all individual scores for all relevant questions.

There is a family of testing formats under the umbrella term Zone Comparison Technique (ZCT), including the Backster, DoDPI, and Utah. These formats are considered to be the most powerful PDD techniques because the relevant questions are very specific and essentially are rewordings of a single question. Testing techniques that employ these types of relevant questions are sometimes called single-issue tests. The original 7-position scoring system was developed on ZCT formats, and this evaluation method is used extensively in the field to render PDD decisions. With single-question tests and 7-position scoring, the grand sum of all individual values can be compared to a threshold, and the decision is based on whether the sum meets or exceeds the cutoffs. The Department of Defense Polygraph Institute standards (DoDPI, 1992) suggest that totals of +6 or greater be labeled No Deception Indicated (NDI); totals between -5 and +5 be labeled No Opinion (NO); and totals of -6 or less are labeled Deception Indicated (DI). Some schools use other thresholds, though DoDPI standards are the most frequently used. Spot scoring, a secondary decision rule, is commonly practiced, but independent research that supports its use is lacking. Spot scoring is addressed later in this paper.

While it is generally agreed within the PDD discipline that the 7-position system produces accurate outcomes, the 7-position system has been criticized by some for imposing an unnecessary subjectivity on the response evaluations (Gordon & Cochetti, 1987; Honts & Driscoll, 1988). Differences among the scores raters assign to specific comparisons may be caused by the requirement that two decisions be made. Raters must first determine which reaction is larger between the relevant and comparison questions, then determine how much larger the reaction is. Identification of the response with the greatest magnitude is usually straightforward, and disagreements among raters are infrequent. However, visually distinguishing proportional differences between responses with analog data is inherently complex, and therefore more vulnerable to individual differences among raters. Consequently, estimates of response magnitude differences, and the associated numerical value, vary more across raters than decisions regarding whether the scores are positive or negative. This explains why examiners evaluating the same test charts will have the same final decisions, but sometimes very different total scores. Decision differences that do occur depend largely on how close to the NO thresholds the scores tend to be.

As an alternative to the 7-position system, some field practitioners use a variant called the 3-position system (Capps & Ansley, 1992b; Shull & Crowe, 1993; Weinstein & Morris, 1990). Its chief attraction is ease of use. The 3-position system entails the assignment of a 1, 0 or -1 to relevant/comparison question pairs, and as such, is a simplified version of the 7-position system. Examiners need to determine which question evoked the greater response, but do not need to estimate how much larger that response is. By reducing the complexity of the required

decisions, variability in totals across raters is expected to be reduced, thereby improving the interrater scoring reliability. This is a hypothetical benefit because no investigator has yet looked into this important question.

There is research supporting the validity of the 3-position scoring system (Capps & Ansley, 1992a; van Herk, 1991). These studies employed both the 3-position and 7-position scoring systems on polygraph charts for which ground truth had been independently established. The investigators concluded that the 3-position system could be as valid as the 7-position. Both Capps et al. and van Herk tried other thresholds, although neither systematically determined the optimum thresholds for the 3-position system, thus leaving the issue to be dealt with heuristically by practitioners. Many users of the 3-position scoring system have defaulted to the standard +/- 6 thresholds of the 7-position scale. When scores fall into the NO region, examiners rescore using the 7-position method in an attempt to obtain a conclusive outcome.

In contrast to previous research, the 3-position scoring system has been approached here as a statistical question: how closely do decisions resulting from 3-position scoring agree with those of 7-position scoring? Seven-position scores assigned by experienced examiners on mock crime data were compared to those of 3-position scores derived from the same data sheets. Cutting scores for the 3-position data were systematically tested to determine which most closely approximated the decisions of the 7-position scoring, and the performance of the two systems were compared.

Method

Three experienced PDD instructors at DoDPI scored 100 sets of PDD recordings as part of a separate study. The PDD recordings had been produced during an earlier laboratory study (see Kircher & Raskin, 1988, for complete details). Half of the 100 sets of PDD recordings were from programmed guilty subjects who committed a mock theft, while the other half were of programmed innocent subjects who were aware that a mock theft took place, but did not participate. The DoDPI scorers were blind to programming, base rates, test questions, or other extra-polygraphic information.

The PDD recordings were scored by the examiners employing the 7-position scale as taught by DoDPI. The strip recordings displayed abdominal and thoracic respiration, skin conductance, cardiovascular, and a plethysmographic tracing. The reviewers scored all five of these physiological channels. In field PDD only three physiological channels are scored (electrodermal activity, cardiovascular activity, and respiration) and only scores assigned to those phenomena were used in the present study. Though the reviewers assigned values to the thoracic and abdominal respiration tracings individually, for the purpose of the present study, only those numbers assigned to the abdominal respiration were considered. Also, the Kircher et al. design entailed the recording of five test charts for each subject, though only the first three were analyzed for this study since this number more closely approximated field practice. It was recognized that additional charts are sometimes used in the field when three recordings are not sufficient for a conclusive call, and that the proportion of definitive decisions would likely be lower in this study because of the cap on the data. However, the researcher adhered to the three-chart rule for the purposes of standardization, and simplicity of decision rules.

In addition to the scorings of the DoDPI examiners, there were blind scorings by both Kircher and Raskin of the University of Utah available. Those data were included in the analysis here, though there were differences in the conditions between the DoDPI and Utah scorers. First, in the Utah scoring system, the scorers always evaluate each relevant question against the comparison question that was presented just before it. The DoDPI scoring method requires scorers to compare to the comparison question with the larger reaction, provided it adjacent to the relevant question. Secondly, though the DoDPI scorers did not know the base rates for deception for the polygraph cases, the experimenters who produced the data would have this information.

Because the DoDPI rules for scoring physiological features with the 3-position scale are identical to those of the 7-position scale except with regard to the range of scores, 3-position scorings in this study were produced by collapsing the values from the 7-position scorings. All +/-3 and +/-2 values for comparisons were reduced to +/-1. Then all values were summed across all questions and all tests, to render a single total value for each examination.

Results

The five scorers produced an average of 66.8% correct, 4.2% incorrect, and 29.0% NO results with the 7-position scale and cutting scores of +/-6. More specifically, for the 250 decisions made on the PDD data from the programmed innocent, there were 177 correct decisions, 6 errors, and 67 NOs. For the 250 decisions on the programmed guilty cases, there were 157 correct, 15 incorrect, and 78 NOs. The overall proportion of correct decisions was significantly greater than chance ($z=5.39$, $p<0.001$). A total of 71.0% of the results were conclusive, and excluding NOs, the raters averaged 94.1% correct decisions. Table 1 shows the individual performance of the blind scorers, and Table 2 is the proportion of agreement among the five scorers.

The performance of the 3-position scale was a function of the thresholds. Table 3 displays the relative accuracy of the 7-position scale, along with the 3-position scale at various symmetrical cutting scores.

Table 1
Accuracy rates of five scorers using the 7-position scale on three sets of polygraph charts from each of 50 programmed innocent and 50 programmed guilty subjects.

	Truthful Subjects				Deceptive Subjects			
	Correct	Error	No Opinion	Total	Correct	Error	No Opinion	Total
Utah scorer 1	34	1	15	50	23	1	26	50
Utah scorer 2	34	1	15	50	26	2	22	50
DoDPI scorer 1	35	2	13	50	36	4	10	50
DoDPI scorer 2	37	1	12	50	35	4	11	50
DoDPI scorer 3	37	1	12	50	37	4	9	50

Table 2
Proportions of agreement on polygraph decisions among five scorers and ground truth employing the 7-position scale and three charts.

	Utah scorer 2	DoDPI scorer 1	DoDPI scorer 2	DoDPI scorer 3	Ground Truth
Utah rater 1	0.84	0.72	0.68	0.72	0.57
Utah rater 2		0.72	0.70	0.76	0.60
DoDPI rater 1			0.74	0.86	0.71
DoDPI rater 2				0.78	0.72
DoDPI rater 3					0.74

Table 3
Number of correct, No Opinion, and incorrect decisions for five scorers of 100 sets of polygraph charts by scoring system and cutting scores (method n = 500 decisions).

		Decisions	
Method & Cutting Score	Correct	No Opinion	Error
7-position (+/-6)	**334**	**145**	**21**
3-position (+/-6)	278	212	10
3-position (+/-5)	312	170	18
3-position (+/-4)	343	132	25
3-position (+/-3)	378	88	34
3-position (+/-2)	403	51	46
3-position (+/-1)	421	21	58

To determine which of the cutting scores of the 3-position scoring system produced results that most closely approximated those of the 7-position system, a goodness of fit test was performed at each of the 3-position cutoffs between +/-1 and +/-6. Of the six thresholds, +/-5, +/-4, and +/-3 were the only cutoffs of the 3-position scale that did not produce proportions of outcomes that were significantly different from the traditional 7-position system. In other words, none of these three cutting scores with the 3-position scale render results significantly different from those of the 7-position system.

The proportions of outcomes from the 3-position scale at thresholds of +/-4 varied less frequently from the proportions of outcomes from the 7-position scale than either the +/-3 or +/-5 (7.7% versus 22.8% and 8.9%, respectively). For simplicity of reporting, the data from the +/-4 cutoffs will be used hereafter. Table 4 places the performance of the 7-position and 3-position scoring systems scores in a form for convenient comparison. Table 5 lists the average scores for 3- and 7-position scoring by programming.

Table 4
Number of correct, incorrect, and No Opinion results from 7-position scoring at +/-6 cutting scores, and 3-position scoring at +/-4 cutting scores. (n=250 decisions per scoring method for each type of programming.)

Programming	Polygraph Decision	7-Position	3-Position
Innocent	No Deception Indicated	177	192
Innocent	No Opinion	67	50
Innocent	Deception Indicated	6	8
Guilty	No Deception Indicated	15	17
Guilty	No Opinion	78	82
Guilty	Deception Indicated	157	151

Table 5
Average total scores for 7- and 3-position scoring systems by programmed guilt and innocence.

	Average Scores	
Programming	7-Postion	3-Postion
Guilty	11.4	7.7
Innocent	-10.4	-6.3

The average Pearson correlation coefficient for total scores among all pairs of blind raters for the 100 sets of strip charts was 0.92 for the 7-position scale. The corresponding average Pearson correlation coefficient for the 3-position scores was 0.90.

An additional decision rule taught by at least two polygraph schools deals with spot scores, or the sums of values of individual relevant questions (Capps & Ansley, 1992b). According to the spot rule, if the total score of any individual relevant question is -3 or lower, the subject is called DI, irrespective of the total for all questions. Total scores between 0 and -2 on any single relevant question result in a NO decision, unless the DI threshold has already been reached for the total score. An NDI result requires a positive value in each spot, and a total of +6 or greater across all spots. The spot rule is applied to multi-issue tests (e.g., screening examinations), multi-facet tests (e.g., did you steal the money, did you know who stole the money, did you spend any of the stolen money), as well as single-issue tests such as those used here (did you do it, did you do it, did you do it).

A post hoc analysis of the 7-position data from the five blind scorers was conducted on the present data for comparative purposes. Table 6 shows how the spot score rule influenced detection efficiency.

Table 6
Average accuracy using the 7-position scale with and without the spot score rule

	Innocent			Guilty		
	Correct	Incorrect	No Opinion	Correct	Incorrect	No Opinion
Spot Rule	57%	10%	33%	79%	4%	17%
No Spot Rule	71%	2%	27%	63%	6%	31%
Difference	-14%	8%	5%	16%	-2%	-14%

When the spot rule was applied to the scores for these cases, there was a substantial gain in the detection of the guilty. Error rates were little affected by the spot rule for this group, and accuracy jumped at nearly the same high rate as the No Opinion percentages decreased. Detection of innocent subjects dropped, however, and there was a substantial increase in error rates for these subjects when the spot rule was used.

There are good reasons to expect the pattern in these results. First, the variability of responding to a single question is greater than across all questions. Therefore, it is not unlikely to have the scores from at least one single question fall into the deceptive range, even for truthful subjects. Second, in contrast to the symmetrical +/-6 cutting scores for whole examinations, the trigger to make a decision of deceptiveness is more sensitive than that for making a decision of truthfulness with the spot rule. For these reasons one would expect a tradeoff in accuracy, with an increase in the detection of

the guilty that corresponded with a similar decrease in the detection of the innocent.

If these data can be generalized to the field, it can be seen that total accuracy when employing the spot rule will be influenced by base rates of deception. The higher the population of deceptive subjects, the greater the accuracy. The use of the spot rule in single-issue examinations where the base rate of innocent examinees is high appears to be ill advised from these data, however.

Discussion

One of the principal purposes of this research was to determine how effective the 3-position scale could be when the optimum cutoffs are applied. The 3-position system thresholds of +/-3, +/-4 and +/-5 produced proportions of outcomes that were not significantly different from those rendered by the orthodox +/-6 thresholds of the 7-position scale. Of the three 3-position thresholds, +/-4 produced the least total variation from the decisions reached by the 7-position scores. Correct, incorrect and NO decisions for the two systems were highly similar with these mock crime data. The present data do not indicate that one of the systems is superior to the other. Indeed, they appear to be interchangeable in this application.

One of the anticipated benefits of the 3-position scoring over the 7-position method was a higher agreement in scoring among scorers, by virtue of the reduction in decision steps. This did not occur as expected, and the reason why the correlation coefficients remained essentially unchanged is unclear. One possible explanation may be the manner in which the 7-position scorings were performed by these scorers. If the scorers tended to use scores that remained in the +/-1 range, the scorers were, by default, employing the 3-position scale though seven positions were available. An inspection of all of the individual values assigned to response comparisons with the 7-position scale revealed that an average of 90.3% of the values were between -1 and +1, leaving only 9.7% for +/-2 and +/-3 values. Similar imbalances in proportions were reported by Capps & Ansley (1992c), who found 79.6% of scores assigned by 11 evaluators for 40 PDD cases were between -1 and +1. Therefore, collapsing the 7-position scores to 3-position scores as was done in this study may have brought about only marginal changes from the original values.

The efficacy of the spot rule with 7-position scoring was disappointing when applied to these single-question examinations. Increases in accuracy for deceptive examinees were offset by an almost equal number of false positives. Considering these findings, and those of Capps & Ansley (1992b) with field cases, the converging evidence indicates that users of the spot rule should be very cautious when scoring these types of tests if the cost attendant to false positive errors is significant. This is because spot scoring appears to increase the likelihood of false positives substantially. A prudent course for practitioners is to report a decision of No Opinion under the combined conditions that total scores are not clearly indicative of deception and a spot score reaches -3 or lower. However, the value of spot scores in multi-issue and multi-facet examinations is likely higher since the relevant questions cover different issues from one another in those tests, and the results of the present analysis would not generalize to those types of cases.

One issue that emerges in many validity studies is the optimum cutting score for the 7-position scale. While the DoDPI standard of +/-6 was used with the 7-position data in the present study, these thresholds are not universally accepted (Capps & Ansley, 1992b), and they have sometimes been called "arbitrary" (Furedy & Heslegrave, 1988). As is apparent in Table 3 regarding the shifting of thresholds, both accuracy and error rates increase or decrease together. While this may seem counterintuitive to some at first, it should be recognized that the establishment of the NO band serves to constrain the proportion of errors in PDD, not maximize the number of correct decisions. Therefore, the issue of what is the "best" threshold is not a simple question. The heart of the question is not which cutting scores produce the most numerous accurate outcomes, but rather, which among the possible cutting scores

produces errors and NO decisions at a rate acceptable to the user. Having an extra-ordinarily wide NO zone, such as +/-30, would virtually eliminate errors, but it would also diminish the value of PDD since only a relative handful of cases would receive these nearly error-free decisions, and the remainder would be No Opinions. Reducing the NO zone to +/- 1 would provide many more correct decisions than the +/-6 thresholds now in common practice, but the concomitant error rate may be unacceptable to many consumers of PDD results. There may not be a best cutting score for all applications, since the cost of errors varies by case. It's worth noting that the prevailing cutting scores for most schools of instruction are a matter of tradition rather than empiricism. This blind spot in the discipline calls for more attention.

References

Backster, C. (1962). Technique tips and polygraph chart interpretation. *Newsletter of the Academy for Scientific Interrogation*, Sep/Oct, 4.

Capps, M.H. & Ansley, N. (1992a). Comparison of two scoring scales. *Polygraph*, 21(1), 39-43.

Capps, M.H. & Ansley, N. (1992b). Analysis of federal polygraph charts by spot and chart total. *Polygraph*, 21(2), 110-131.

Capps, M.H. & Ansley, N. (1992c). Numerical scoring of polygraph charts: What examiners really do. *Polygraph*, 21(4), 264-320.

Department of Defense Polygraph Institute. (1992). *Zone comparison test.* Fort McClellan, Alabama.

Furedy, J.J. & Heslegrave, R.J. (1988). Validity of the lie detector: A psychophysiological perspective. *Criminal Justice and Behavior*, 15(2), 219-246.

Gordon, N.J. & Cochetti, P.M. (1987). The horizontal scoring system. *Polygraph*, 16(2), 116-125.

Hathaway, S.R. & Hanscom, C.B. (1958). The statistical evaluation of polygraph records. In V.A. Leonard (Ed.), *Academy lectures on lie detection, 2*, 111-121. Springfield, IL: Charles C. Thomas.

Honts, C.R. & Driscoll, L.N. (1988). A field validity study of the rank order scoring system (ROSS) in multiple issue control question tests. *Polygraph*, 17, 1-16.

Kircher, J.C., & Raskin, D.C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, 73(2), 291-302.

Lee, C.D. (1943). *Instruction manual for the Berkeley psychograph*, CA: C.D. Lee & Sons.

Lee, C.D. (1953). *The instrumental detection of deception - The lie test.* Springfield, IL: Charles C. Thomas.

Lykken, D.T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*(6), 385-388.

Lykken, D.T. (1979). The detection of deception. *Psychological Bulletin*, 86(1), 47-53.

Lykken, D.T. (1981). *A tremor in the blood: Uses and abuses of the lie detector.* New York: McGraw-Hill.

Lykken, D.T. (1988). Detection of guilty knowledge: A comment on Forman and McCauley. *Journal of Applied Psychology,* 73(2), 303-304.

Shull, K.W. & Crowe, M. (1993). *Effects of two methods of comparing relevant and control questions on the accuracy of psychophysiological detection of deception. Department of Defense Polygraph Institute,* DoDPI-R-0002.

van Herk, M. (1991). Numerical evaluation: Seven point scale +/- 6 and possible alternatives; A discussion. *Polygraph,* 20(2), 70-79.

Weinstein, D., & Morris, P.J. (1990) *A comparative investigation of the reliability between differing polygraph scoring techniques.* Unpublished manuscript.

Winter, J.E. (1936). A comparison of the cardio-pneumo-psychograph and association methods in the detection of lying in cases of theft among college students. *Journal of Applied Psychology,* 20, 243-248.

Criterion Development and Validity of the CQT in Field Application

Charles R. Honts[1]

Abstract

A field study of the control question test (CQT) for the detection of deception was conducted. Data from the files of 41 criminal cases were examined for confirming information and were rated by two evaluators on the strength of the confirming information. Those ratings were found to be highly reliable, r = .94. Thirty-two of the cases were found to have some independent confirmation. Numerical scoring and decisions from the original examiners and independent evaluation were analyzed. The results indicated that the CQT was a highly valid discriminator. Excluding inconclusives, the decisions of the original examiners were correct 96% of the time, and the independent evaluations were 93% correct. These results suggest that criteria other than confession can be developed and used reliably. In addition, the validity of the CQT in real-world settings was supported.

Keywords: Comparison question, control question, field study, polygraph, validity.

Psychophysiological credibility assessment (also known as the psychophysiological detection of deception, lie detection, or polygraph) is an important application of psychology to the real world (Honts, 1994). The use of polygraph examinations for forensic investigation is widespread in law enforcement in the United States and Canada (Honts, 1994). Moreover, the results of polygraph examinations have been gaining acceptance as evidence in American courts of law (Honts & Quick, 1995). Worldwide, the control question test (CQT; Raskin, 1989) is the most commonly used polygraph test in law enforcement (Raskin, Honts, & Kircher, 1997).

The CQT assesses a person's credibility by looking for a differential reaction between two types of questions. The first type of question is known as a relevant question. Relevant questions are direct accusatory questions that address the issue under investigation (e.g., Did you shoot Joy Doe?). Control questions are ambiguous questions to which the examinee is maneuvered into answering in the negative (i.e., Before 1994, did you ever do anything that was dishonest, illegal, or immoral?). The rationale of the CQT is that guilty persons will have larger physiological responses to those questions they answer with a lie than to the relatively unimportant control questions. Innocent persons are expected to have larger reactions to the control questions (with responses that are assumed to either be untrue or, at least, uncertain) than to the truthfully answered relevant questions. Unfortunately, the actual

[1]Department of Psychology, Boise State University.

The research reported here was supported by a contract from the Canadian Police Research Center, Science Branch, contract number M9010-3-2219/01-ST. The author wishes to acknowledge and thank S/Sgt. John W. Kaster of the Polygraph Training Unit of the Canadian Police College for his considerable effort, assistance, and support throughout this project. The author also wishes to thank Mary Devitt for her help in the initial preparation of this article and David Raskin and John Kircher for their thoughts and comments.

The comments and conclusions in this report are those of the author. They do not necessarily reflect the official position or policy of the Canadian Government, the Royal Canadian Mounted Police, the Canadian Police College, or the Polygraph Training Unit.

Reprinted with the permission of the author and *The Journal of General Psychology.* The citation for the original publication is *Journal of General Psychology* 1996, 123(4), 309-324. Address correspondence to Dr. Charles R. Honts, Department of Psychology, Boise State University, 1910 University Drive, Boise, ID 83725; e-mail: honts@truth.idbsu.edu.

validity of the CQT has been, and continues to be, the subject of a polemic debate in the scientific community (see, e.g., Furedy, 1993, vs. Honts, Kircher, & Raskin, 1995; see also the review in Honts & Perry, 1992).

Two recent comprehensive reviews arrived at very different conclusions about the accuracy of the CQT (Iacono & Patrick, 1988; Raskin, 1989). Iacono and Patrick concluded that "the best defense one can offer for the continued use of the CQT is that its accuracy is indeterminate" (p. 233). Raskin concluded, "The voluminous scientific literature indicates that they [CQTs] can be highly accurate when properly employed in appropriate circumstances" (p. 290). These strikingly different conclusions are based on (a) differing opinions regarding the value of laboratory experiments and (b) which field studies are considered to have adequate methodology.

Critics of polygraph tests (e.g., Iacono & Patrick, 1988; Kleinmuntz & Szucko, 1982; Lykken, 1981) generally dismiss the results of all laboratory simulation studies as useless for estimating field accuracy. They argue that the qualitative context produced by the threat of criminal sanctions in the real world cannot be simulated in the laboratory. Others (Kircher, Horowitz, & Raskin, 1988; Kircher, Raskin, Honts, & Horowitz, 1994; Raskin, 1989) have suggested that the differences between laboratory and field settings may not be that great. They have argued that if simulation studies use representative populations and realistic polygraph practices, and include some motivation to deceive, then they can provide useful information for estimating field accuracy. Kircher *et al.* (1988) conducted a meta-analysis on 14 laboratory studies of the CQT. They reported an average accuracy of 87% for the five studies (Ginton, Netzer, Elaad, & Ben-Shakhar, 1982; Kircher & Raskin, 1982; Podlesny & Raskin, 1978; Raskin & Hare, 1978; Rovner, Raskin, & Kircher, 1979) that they rated as most ecologically valid. The four studies (Barland & Raskin, 1975; Bradley & Ainsworth, 1984; Bradley & Janisse, 1981; Szucko & Kleinmuntz, 1981) that they rated lowest in ecological validity produced an average accuracy rate of only 73%.

Given that high-quality laboratory studies have demonstrated that the CQT can be highly accurate, the next step is to examine its use in the field. The primary concern is one of generalization. It may be that laboratory simulations are unable to adequately model the field phenomenon. Obviously, it is not possible to bring the kind of sanctions faced by a criminal suspect into the laboratory. It may be that psychophysiological credibility assessment in the laboratory and in the field are qualitatively different phenomena. Although recent research has suggested that this is probably not the case (Kircher *et al.*, 1994), only field studies can provide the evidence needed to demonstrate generalizability from laboratory to field settings.

Accuracy estimates for the CQT, based on field studies, have varied wildly, ranging from chance to near perfection. In considerations of the field studies, the scientific arguments have generally centered on the methodology used in the various experiments. Honts and Quick (1995) recently reviewed the literature concerning field validity studies of the CQT. They noted that the adequacy of field studies has generally been evaluated using the following four factors: subjects, evaluation method, sampling strategy, and criterion development. It seems that members of the scientific community (Honts & Quick, 1995; Iacono & Patrick, 1988; Lykken, 1981; Raskin, Honts, & Kircher, 1997) generally agree that the following are necessary for a useful field study of psychophysiological credibility assessment.

1. If the primary target for generalization is the application of polygraph testing in law enforcement, then the subjects of field validity studies should be suspects in real-life criminal cases. Questions about the validity of the CQT with victims and other types of subjects are of interest, but such questions require separate examination.

2. Evaluations should be based on methods that rely only on the physiological data. Moreover, the evaluations should be conducted by persons trained and experienced in using only physiological data to evaluate credibility. Those evaluations should use scoring techniques that are representative of those used in the field.

3. The sampling of cases should be according to an acceptable scientific basis. Cases must not be selected on the basis of the quality of the charts or on the accuracy of the outcome of the original examiner. Patrick and Iacono (1991) have argued that an exhaustive sampling strategy is likely to produce the minimum amount of sampling error. Exhaustive sampling refers to sampling in which all available cases from a specified time period are included in the study.

4. A criterion that is independent of the polygraph test must be developed for who is innocent and who is guilty. Generally, confessions have been considered to be the only acceptable criterion. Unfortunately, the use of confession as a criterion introduces a number of problems of sampling bias, which, in turn, raise questions about the usefulness of confession studies (Patrick & Iacono, 1991). In addition, confessions are sometimes false (Kassin & Kiechel, 1996). When confessions are used as a criterion in polygraph field studies, their reliability is enhanced if additional evidence is later found to support the validity of the confession (e.g.: The subject confesses to having stolen the missing item and tells investigators where the item is hidden. The investigators subsequently find the missing item where the suspect said it was hidden). However, for now, confession-based criteria appear to be the best available criteria, especially if the confession is supported by evidence. It is clear that developmental work is needed to determine whether viable alternatives to confession criteria are possible and useful.

In 1983, the United States Congress's Office of Technology Assessment (OTA, 1983) conducted an extensive review of the available polygraph field studies. They found 10 studies that met their minimal standards and reported an accuracy rate of 90% for criterion-guilty subjects and an accuracy rate of 80% for criterion-innocent subjects. Lamentably, none of those studies adequately satisfies all of the criteria specified above.

Since the OTA study, three new field studies (Honts & Raskin, 1988; Patrick & Iacono, 1991; Raskin *et al.*, 1988) have been reported. Honts and Perry (1992) argued that all of these studies appear to satisfy the methodological criteria mentioned above. (See Table 1 for the classification results for the independent evaluators and original examiners in those three studies.) Across those three studies, when inconclusive outcomes were excluded, the independent evaluators correctly classified 78% of the criterion-innocent subjects and 98% of the criterion-guilty subjects. It is interesting to note that the independent evaluations in the studies by Honts and Raskin (1988) and Raskin *et al.* (1988) were at least as accurate as those of the original examiners. However, the independent evaluators in the Patrick and Iacono study of Royal Canadian Mounted Police (RCMP) were much less accurate with innocent subjects than the original examiners were. The original examiners correctly classified 90% of the innocent subjects, whereas the independent evaluators' decisions were only 55% correct. The reasons for the differences between those studies are not apparent and deserve consideration.

Table I
Percent Correct Decisions in Three High-Quality Field Studies

	Original examiners		Independent evaluators	
Study	Innocent	Guilty	Innocent	Guilty
Honts & Raskin (1988)	100	92	100	92
Raskin *et al.* (1988)	96	95	89	100
Patrick & Iacono (1991)	90	100	55	98

Note: Inconclusive outcomes were excluded from these calculations.

The Patrick and Iacono (1991) study may have suffered from a problem known as criterion contamination (Muchinsky, 1993). That is, they may have been measuring something other than just the validity of the CQT. For example, recent laboratory (Barland, Honts, & Barger, 1989) and field (Raskin *et al.*, 1988) research has suggested that the CQT does not have very good specificity. That is, the CQT can determine whether the subject is attempting deception regarding some issue, but it may not be very good at determining which issue is being responded to deceptively when more than one issue is addressed.

Honts and Raskin (1988) and Raskin *et al.* (1988) directly addressed this problem methodologically by explicitly considering it in the rules for subject classification and presenting accuracy rates for single relevant issues within examinations (Raskin *et al.*). There is no indication that Patrick and Iacono (1991) took such issues into account.

Another issue might contribute to criterion contamination. RCMP polygraph examinations usually address the most serious level of involvement under investigation. If the suspect being tested was not guilty of the most serious level of involvement but was involved in the crime, an issue arises about how such a subject should be classified. Consider the following scenario, a case from the Patrick and Iacono (1991) study (S/Sgt. J. Kaster, personal communication, 1991). A diamond ring was stolen. Suspect A took a polygraph test wherein the relevant questions were of the form, "Did you steal the diamond ring?" The suspect failed the polygraph test and subsequently confessed that although he did not steal the diamond ring, his brother did, and he (Suspect A) sold the stolen property. Is Suspect A truthful or deceptive with regard to the polygraph examination? The position taken in the present study was that Suspect A would have been considered deceptive, because his intention was to deceive the polygraph examiner. However, Patrick and Iacono considered the subject to be truthful (J. Kaster, personal communication, 1991), suggesting that criterion contamination may have been a serious problem with their study. It would be useful to examine the original case files from this study, to determine the exact extent of any such problem. Unfortunately, the files were not kept by the investigators (C. Patrick, personal communication, 1996), and the RCMP detachments to whom they belong will not release them for further study (J. Kaster, personal communication, 1991).

The present study was undertaken to address some of the problems associated with conducting field studies in the detection of deception and to obtain another estimate of the validity of the CQT as used by the RCMP. I obtained complete files from cases in which at least one polygraph examination had been conducted from Canadian law enforcement with the cooperation of the Canadian Police College. The materials in those case files were then evaluated for confirmation information independent of the polygraph examination outcome. Two independent evaluators provided independent ratings of strength of confirmation. The polygraph data from the cases was then independently evaluated. My expectation was that when the criterion contamination issues were addressed, the accuracy rates of our independent evaluators would be similar to those of the original examiners. In addition, I hoped to provide some initial information concerning the use of criteria other than confessions in conducting field studies of the detection of deception.

Method

Obtained Cases

The sampling of cases occurred in two waves. The obtained cases from the first wave represented an exhaustive sample (*i.e.*, all cases) of 1 year's cases (23 cases) from a single polygraph examiner. Those cases included 29 polygraph tests (in some cases, more than one suspect was given a polygraph).

The second wave of data collection provided materials from 12 polygraph examinations. Those cases were also obtained from a single examiner (different from the one used in the first collection of data). Those cases represented an exhaustive sample of all of that examiner's cases that contained any confirmatory information other than the polygraph result. This sample also represented 1 year's work for the examiner.

Strength of Confirmation Ratings

Two evaluators rated the strength of confirmation provided by the respective case files from the first wave of data collection. One of the evaluators was a highly experienced (20 years) senior police investigator and polygraph examiner. The second evaluator had a Ph.D. in psychology and had been a polygraph examiner and investigator for 16 years. The evaluators were given the complete case files and worked independently. They used a 7-point scale that ranged from 1, no confirmation, to 7, very strong confirmation.[1] Category 1 indicated no confirming information other than the polygraph test.

The two ratings were correlated and found to be very similar, $r = .94$, $p < .0001$. This result indicates that the procedures used to estimate the strength of confirmation were very reliable. Given the strong results in the first data wave, only one assessment of the strength of confirmation was obtained in the second wave. For pragmatic reasons, it was decided that the psychologist would make the strength of confirmation ratings in the second wave.

Participant Demographics

Data from 41 suspects were examined. Ten suspects were female. Suspects' ages ranged from 15 to 62 years, with a median age of 24 years and a mean age of 26.5 years, SD = 11.34. Their number of years of education ranged from 7 to 14 years, with a mean of 10.22 years, SD = 1.87.

Criteria for Cell Inclusion

The data from both waves of data collection were combined for analysis, and four categories of confirmation were created from the strength of confirmation ratings. Criteria for the initial inclusion of cases in the various categories of confirmation were as follows:

Strong confirmation. The case must have received a strength of confirmation of 5 or greater. Further, to be included in the strong confirmation condition, the perpetrator must also have confessed to the crime under investigation.

Moderate confirmation. The case must have been rated at a strength of confirmation of at least 5. However, no confession was necessary.

Weak confirmation. Cases that have some confirmation information but ratings of strength of confirmation of less than 5.

No confirmation. Cases without confirmatory information.

Based on the above criteria, 7 suspects were selected into the no confirmation category, 11 suspects were selected into the weak confirmation category, 10 suspects were selected into the moderate confirmation category, and 13 participants fit into the strong confirmation category. Of the 7 suspects put into the no confirmation category, 6 were classified by the original examiner as truthful, and the one remaining examination produced an inconclusive outcome. For the 11 suspects selected into the weak confirmation category, the available evidence indicated that 3 of them were truthful, whereas the data for 6 others contained information indicating that they had been deceptive. The two evaluators disagreed on the valance of the evidence regarding the remaining 2 suspects in this category, who were dropped from subsequent analyses. The evaluators agreed regarding all other suspects. For the 10 suspects in the moderate confirmation category, the evidence in the case files indicated that 8 were deceptive and 2 were truthful. For the 13 suspects in the strong confirmation category, the evidence and

[1] The following examples were given to the evaluators to guide their ratings: "A confession followed by the development of physical evidence based on information in confession (a person confesses to a murder and describes where the body is buried, the body is found hidden as described) would deserve a rating of 7. A recantation of a victim without any additional evidence would deserve an intermediate rating of 3, 4, or 5, depending on the context of the recantation and the quality of the recantation. A statement by a single informant without any supporting physical evidence would be rated as a very weak confirmation, a 1 or 2 on the scale." (Honts, 1991, p. 12).

a confession indicated that 6 suspects were truthful (there was a confession of guilt plus evidence provided by another suspect), and 7 were deceptive.

Numerical Evaluations

The original examiners and an independent evaluator performed numerical evaluations according to the techniques taught at the Canadian Police College (CPC). A second set of independent numerical evaluations was made on the data collected in the second wave. The CPC numerical scoring techniques were based on the numerical scoring system developed and validated at the University of Utah (Kircher & Raskin, 1988). That numerical scoring system has been shown to be both highly reliable and highly valid in a number of laboratory and field studies (for a review, see Raskin, 1989).

In the CPC numerical scoring system the physiological responses of the suspect are evaluated at each relevant and control question pair for the presence of the following criteria: respiration--decrease in amplitude, slowing of rate, and increase in baseline; electrodermal response--increase in amplitude, increase in duration, and increase in complexity (number of phasic responses); cardiovascular response--increase in ampli-tude of the slow wave and increase in duration of the slow-wave response. Additional details regarding these criteria can be found in Kircher and Raskin (1988).

At each relevant and control question pair, each physiological system was evaluated independently. At each comparison point, a score was assigned on a 7-point scale that ranged between -3 and +3. If a criterion-defined response to the relevant question in the pair was stronger, a negative score was assigned. If a criterion-defined response to the control question in the pair was stronger, a positive score was assigned. Magnitude of the assigned score reflected a judgment about the degree of difference between the magnitude of the response to relevant and control questions. Equivalent responses to both questions, including no response to both questions, resulted in a score of 0. After all relevant and control question pairs had been scored, all of the scores were summed. The total numerical score was then evaluated to make a decision regarding the participant's credibility. Total numerical scores greater than +5 resulted in a decision of truthful. Total numerical scores less than -5 resulted in a decision of deceptive. Total numerical scores between -5 and +5, inclusive, resulted in an inconclusive outcome.

Results

Reliability

Total numerical scores generated by the original examiners and the independent evaluator were available for all 41 suspects. The scores were correlated. The resulting correlation was significant, $r(39) = .91$, $p < .0001$. The total numerical scores from the second wave of data collection from the original examiners and the two independent evaluators were also correlated (see Table 2). The results indicated that the numerical scoring system developed at the University of Utah and taught at the Canadian Police College's Polygraph Training Unit is highly reliable. These results are comparable to the results obtained in highly controlled laboratory studies (Kircher & Raskin, 1988).

Table 2
Interrater Reliability Coefficients for the Three Numerical Evaluations

Evaluator	Independent 1	Independent 2
Original examiner	.91	.96
Independent 2	.92	

Numerical Scores

The numerical scores of the original examiners and the independent evaluators who scored all of the data were examined with a guilt (innocent, guilty) by level of confirmation (high, medium, low) ANOVA (see Table 3, for the means). Participants for whom there was no confirmatory information were not included in this analysis. The ANOVA revealed a main effect of guilt in the scores of the original examiners, $F(1, 26) = 45.68$, $p < .001$, and in the scores of the independent evaluator, $F(1, 26) = 30.55$, $p < .001$. In neither data set were the main effects of level of confirmation (for the original examiner, $F[2, 26] = 1.29$, ns; for the independent evaluator, $F[2, 26] = 0.97$, ns) or the interaction involving level of confirmation significant--(original examiner, $F[2, 26] = 0.86$, ns; independent evaluator, $F[2, 26] = 2.58$, ns).

Because there were no significant level-of-confirmation effects, the data were collapsed across levels of confirmation for the initial validity analysis. The combined mean numerical scores generated by the original examiners and the independent evaluator for guilty and innocent suspects are shown in Table 3. The validity of the numerical scores was tested in two ways. Initially, ANOVA was used to test for differences between the numerical scores assigned to innocent and guilty suspects. Then, correlations were calculated between the numerical scores and the guilt criterion, to index the discriminative power of the numerical scores (Kircher *et al.*, 1988). The ANOVAs revealed significant main effects for guilt--for the original examiners,

Table 3
Mean Numerical Scores and Standard Deviations, by Guilt, Level of Confirmation, and Evaluator

Evaluator/ Guilt status	Level of confirmation Low	Medium	High	Combined	*r*
Original examiners					
Innocent					
M	8.67	9.00	8.00	8.36	
SD	1.53	2.82	4.52	3.41	
n	3	2	6	11	
					.76
Guilty					
M	-5.00	-6.62	-13.28	-8.38	
SD	10.97	7.58	4.15	8.27	
n	6	8	7	21	
Independent evaluator					
Innocent					
M	-4.67	6.00	8.17	4.27	
SD	1.52	5.66	5.27	7.16	
n	3	2	6	11	
					.71
Guilty					
M	-11.23	-12.00	-15.00	-12.81	
SD	12.13	9.38	5.45	8.89	
n	6	8	7	21	

$F(1, 31) = 40.88$, $p < .001$, and for the independent evaluator, $F(1, 31) = 30.18$, $p < .001$). The validity correlations were very strong for both the original examiners and the independent evaluator, accounting for 58% and 50% of the criterion variance, respectively. That was very good performance, on a par with the strongest results reported in high-quality laboratory studies (Kircher *et al.*, 1988).

To provide a reasonable comparison with other field studies of the detection of deception, I performed separate analyses on only those examinations that had been confirmed by confession. The means from those analyses are shown in Table 4. An ANOVA revealed a significant main effect for guilt in the numerical scores of both the original examiners, $F(1, 11) = 78.38$, $p < .001$, and the independent evaluator, $F(1, 11) = 60.20$, $p < .001$. The correlations assessing the discriminative power of the numerical scores were strong in the subset, with values of .94 and .92 for the original examiners and the independent evaluator, respectively.

Table 4
Mean Numerical Scores, Standard Deviations, and Detection Efficiency *r* Values for Innocent and Guilty Participants Confirmed by Confession

Evaluator	Innocent	Guilty
Original examiner		
M	8.00	-13.29
SD	4.52	4.15
n	6	7
Detection $r = .94$		
Independent evaluator		
M	8.17	-15.00
SD	5.27	5.45
n	6	7
Detection $r = .92$.		

Decisions

Decisions were derived from the numerical scores with the standard field decision rule described earlier. As with the numerical scores, the decisions of the original examiners and the independent evaluator were initially analyzed without reference to level of confirmation. The original examiners correctly classified 81.8% (9 out of 11) of the innocent suspects and called 18.2% (2 out of 11) of them inconclusive. No innocent suspects were incorrectly classified. The original examiners correctly classified 71.4% (15 out of 21) of the guilty suspects, incorrectly classified 4.8% (1 out of 21) of them, and called 23.8% (5 out of 21) of them inconclusive. Excluding inconclusive outcomes, 96% of the original examiners' decisions were correct.

The independent evaluator correctly classified 54.5% (6 out of 11) of the innocent suspects, incorrectly classified 9.1% (1 out of 11) of them, and called 36.4% (4 out of 11) of them inconclusive. In classifying guilty suspects, the independent evaluator correctly classified 90.5% (19 out of 21) of the participants, incorrectly classified 4.8% (1 out

of 21), and called 4.8% (1 out of 21) of the suspects inconclusive. Excluding inconclusive outcomes, 93% of the independent evaluator's decisions were correct.

The power of the original examiners' and the independent evaluator's decisions in discriminating between truth tellers and deceivers was assessed by coding the guilt criterion (0,1) and the decisions (1 = truthful, 2 = inconclusive, 3 = deceptive) and then correlating the resulting data vectors. This analysis produced a detection efficiency coefficient, useful in making comparisons of discriminative power across studies (Kircher *et al.*, 1988). The detection efficiency coefficient for the original examiners was 0.81. The detection efficiency coefficient for the independent evaluator was 0.76. These are very strong detection efficiency coefficients, indicating that the conclusions reached by both the original examiners and the independent evaluator discriminated well between truth tellers and deceivers.

To provide measures of decision accuracy that could be used in comparisons with other field studies of the detection of deception, I analyzed the decision accuracy of the original examiners and the independent evaluator for only those cases confirmed with a confession. With innocent suspects, the original examiners correctly classified 66.7% (4 out of 6) and called 33.3% (2 out of 6) of the suspects inconclusive. No innocent suspects were incorrectly classified. With guilty suspects, the original examiners correctly classified 100% (7 out of 7) of the suspects. There were no inconclusive or incorrect outcomes. Excluding inconclusives, 100% of the original examiners' decisions were correct. With innocent suspects, the independent evaluator correctly classified 83.3% (5 out of 6) and called 16.7% (1/6) inconclusive. There were no incorrect classifications of innocent suspects. With guilty suspects, the independent evaluator correctly classified 100% (7 out of 7) of the suspects. There were no incorrect or inconclusive outcomes. Excluding inconclusive outcomes, 100% of the independent evaluator's decisions were correct.

I also calculated detection efficiency coefficients for the original examiners' and the independent evaluator's decisions. The detection efficiency coefficient for the original examiners was 0.93, $p < .001$. The detection efficiency coefficient for the independent evaluator was 0.96, $p < .001$.

In making their decisions, the original examiners in this study did not always follow the standard decision rule described above. In four cases involving 2 innocent and 2 guilty suspects, even though the total numerical score for the case was inconclusive, the examiner went ahead and rendered a decision. In all four cases, those decisions were correct. In all cases for which there was some confirmation, the original examiners' decisions were 100% correct regarding the innocent suspects; with the guilty suspects, they correctly classified 81% (17 out of 21), incorrectly classified 4.8% (1 out of 21), and called 14.3% (3 out of 21) inconclusive. Excluding inconclusives, 97% of the original examiners' decisions were correct. When only the 13 cases confirmed by confession were considered, the field calls of the original examiners were 100% correct regarding all suspects. There were no incorrect or inconclusive outcomes. For all cases with some confirmation, the detection efficiency coefficient for the field decisions was 0.89, $p < .001$. For only the cases with the strongest level of confirmation, the detection efficiency coefficient was 1.00.

Discussion

The present study was successful in achieving a number of goals. A field study was conducted using actual law-enforcement case files that met scientific standards: Exhaustive sampling was used without reference to chart quality or the original examiners' decision. Independent evaluations based only on the physiological data were made by an experienced evaluator who used a standard numerical scoring method. Confessions were used to provide a group of cases with the strongest possible level of confirmation. Explicit operational definitions were used, to avoid criterion contamination. The study provided data suggesting that there may be alternatives to the confession criterion.

The study yielded four interesting findings. First, I obtained strong evidence that

the numerical scoring system is very reliable. Interrater correlation coefficients were all above .9, indicating excellent reliability for the numerical scoring system. This finding is not surprising, because the numerical scoring technique taught at the CPC Polygraph Training Unit is based on the numerical scoring system developed at the University of Utah (Kircher & Raskin, 1988; Podlesny & Raskin, 1978; Raskin & Hare, 1978), a system that has consistently been found to be highly reliable (Raskin, 1986). However, the present results demonstrate that the reliability of the technique extends to field settings in Canada. This finding is similar to, and complements the results of Raskin *et al.* (1988) with the United States Secret Service.

The second finding of interest concerns the validity of polygraph tests conducted in the field by examiners using the CQT. The results suggest that this technique is very accurate in discriminating between truth tellers and deceivers in field settings. Both the numerical scores and the decisions based on them provided strong validity coefficients. The results in this field study are as strong as the best results seen in high-quality laboratory studies and other adequately controlled field studies; they suggest that the CQT is a highly valid tool for use by law enforcement in the field.

It is interesting to consider why the results of this study contrast so strikingly with the results of the study reported by Patrick and Iacono (1991). In that study, there were many more false positive errors than there were in the present study. These differences may be related to differences between the operational definitions of guilty and innocence used, or they may be related to criterion contamination. Without examining the actual raw data from the Patrick and Iacono study, it is not possible to determine the exact nature of the methodological differences. Unfortunately, the data from the Patrick and Iacono study are not available (C. Patrick, personal communication, 1996).

The other major difference between this study and the Patrick and Iacono (1991) study concern the difference between the accuracy of the original examiners and the independent evaluators. In the Patrick and Iacono study, there was a tremendous loss of accuracy between original examiners and the independent evaluators, with the false positive rate being many times higher for the independent evaluations. In the present study, the independent evaluations were only slightly less accurate than the decisions made by the original examiners. Moreover, when only the cases with the highest level of confirmation were considered, the independent evaluations were slightly more accurate than the evaluations by the original examiners. This results is in sharp contrast to Patrick and Iacono's results, but it is consistent with the two other field studies that have used similar methods. In the field studies reported by Honts and Raskin (1988) and Raskin *et al.* (1988), the independent evaluations were nearly as accurate as the evaluations of the original examiners (see Table 1). It is not possible to know what happened in the Patrick and Iacono study to cause the independent evaluations to be of such low accuracy, but in comparison with other high-quality field studies, the Patrick and Iacono study can be seen to be an outlier.

The third interesting finding of this study concerns the cases for which the numerical scores supported only an inconclusive outcome and the original examiners chose to override the scoring rules to render a decision. In this study, they were always correct in such calls. Admittedly, this finding covers only four cases and is of limited generalizability, but it poses some interesting questions. What is it about those four cases that led the examiners to break the rules? Can the information used by the examiners to make the decision to break the rules be objectified and used in a systematic way? These questions are of interest and deserve study.

The fourth interesting finding concerns the strength-of-confirmation ratings. The results of this study suggest that the process of rating the strength of confirmation may be a useful way to approach criterion development in field studies of the detection of deception. The present approach to providing such ratings was highly reliable. Although that does not indicate validity for the approach, it is an important and necessary first step in that direction.

I found no significant differences in numerical scores across the levels of confirmation. This finding must be qualified by the fact that the present study is of relatively low power to find such effects, and a null finding under such circumstances should not be given strong weight. Nevertheless, it is interesting to note that the means for the medium and high levels of confirmation were almost identical (e.g., for the independent evaluator with innocent suspects, the mean for medium was 6.0 and for high was 8.17; for the independent evaluator with guilty suspects, the mean for medium was -12 and for high was -15). This finding suggests that if there are level-of-confirmation differences, they may be of small magnitude. It may therefore be possible to combine such categories without a loss of accuracy in the criterion. This would be of great benefit: It would make the acquisition of data from the field much easier and might help to avoid some of the sampling problems that have plagued field studies in this area.

References

Barland, G.H., Honts, C.R., & Barger, S.D. (1989). The validity of detection of deception for multiple issues. *Psychophysiology*, 26, S13 (Abstract).

Barland, G.H., & Raskin, D.C. (1975). An evaluation of field techniques in detection of deception. *Psychophysiology*, 12, 321-330.

Bradley, M.T., & Ainsworth, D. (1984). Alcohol and psychophysiological detection of deception. *Psychophysiology*, 21, 63-71.

Bradley, M.T., & Janisse, M.T. (1981). Accuracy demonstrations, threat, and the detection of deception: Cardiovascular, electrodermal, and pupillary measures. *Psychophysiology*, 18, 307-315.

Furedy, J.J. (1993). The "control" question "test" (CQT) polygrapher's dilemma: Logico-ethical considerations for psychophysiological practitioners and researchers. *International Journal of Psychophysiology*, 15, 263-267.

Ginton, A., Netzer, D., 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.

Honts, C.R. (1991). *Task I Report: Field validity study of Canadian Police College Polygraph Technique. Contract No. M9010-1-F107/01-ST.* Progress report submitted to the Canadian Police Research Center.

Honts, C.R. (1994). The psychophysiological detection of deception. *Current Directions in Psychological Science*, 3, 77-82.

Honts, C.R., Kircher, J.C., & Raskin, D.C. (1995). Polygrapher's dilemma or psychologist's chimaera: A reply to Furedy's logico-ethical considerations for psychophysiological practitioners and researchers. *International Journal of Psychophysiology*, 20, 199-207.

Honts, C.R., & Perry, M.V. (1992). Polygraph admissibility: Changes and challenges. *Law and Human Behavior*, 16, 357-379.

Honts, C.R., & Quick, B.D. (1995). The polygraph in 1995: Progress in science and the law. *North Dakota Law Review*, 71, 987-1020.

Honts, C.R., & Raskin, D.C. (1988). A field study of the validity of the directed lie control question. *Journal of Police Science and Administration*, 16, 56-61.

Iacono, W.G., & Patrick, C.J. (1988). Assessing deception: Polygraph techniques. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 205-233). New York: Guilford.

Kassin, S.M., & Kiechel, K.L. (1996). The social psychology of false confessions: Compliance, internalization, and confabulation. *Psychological Science*, 7, 125-128.

Kircher, J.C., Horowitz, S.W., & Raskin, D.C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior*, 12, 79-90.

Kircher, J.C., & Raskin, D.C. (1982). Cross-validation of a computerized diagnostic procedure for detection of deception. *Psychophysiology*, 20, 568-569. (Abstract)

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

Kircher, J.C., Raskin, D.C., Honts, C.R., & Horowitz, S.W. (October, 1994). *Generalizability of statistical classifers for the detection of deception.* Paper presented at the annual meeting of the Society for Psychophysiological Research, Atlanta, Georgia.

Kleinmuntz, B., & Szucko, J.J. (1982). On the fallibility of lie detection. *Law and Society Review*, 17, 85-104.

Lykken, D.T. (1981). *Tremor in the blood: Uses and abuses of the lie detector.* New York: McGraw-Hill.

Munchinsky, P.M. (1993). *Psychology applied to work: An introduction to industrial and organizational psychology* (4th ed.). Pacific Grove, CA: Brooks/Cole.

Office of Technology Assessment (1983). *Scientific validity of polygraph testing: A research review and evaluation--A technical memorandum (OTA-TM-H-15).* Washington, D.C.: U.S. Government Printing Office.

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, 344-358.

Raskin, D.C. (1986). The polygraph in 1986: Scientific, professional and legal issues surrounding application and acceptance of polygraph evidence. *Utah Law Review*, 1986, 29-74.

Raskin, D.C. (1989). Polygraph techniques for the detection of deception. In D.C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 247-296). New York: Springer.

Raskin, D.C., & Hare, R.D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology*, 15, 126-136.

Raskin, D.C., Honts, C.R., & Kircher, J.C. (1997). The scientific status of research on polygraph techniques: The case for polygraph tests. In D.L. Faigman, D. Kaye, M.J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony*, 565-582

Raskin, D.C., Kircher, J.C., Honts, C.R., & Horowitz, S.W. (1988). *A study of the validity of polygraph examinations in criminal investigations.* Final report to the National Institute of Justice, Salt Lake City, University of Utah, Department of Psychology. (Grant No. 85-IJ-CX-0400).

Rovner, L.I., Raskin, D.C., & Kircher, J.A. (1979). Effects of information and practice on detection of deception. *Psychophysiology*, 16, 198. (Abstract)

Szucko, J.J., & Kleinmuntz, D. (1981). Statistical versus clinical lie detection. *American Psychologist*, 36, 488-496.

Ethical Issues in Sex Therapy and Research Volume 2

Edited by

William H. Masters, M.D., Virginia E. Johnson, D.Sc (Hon.), Robert C. Kolodny, M.D., and Sarah M. Weems, M.A.

Boston: Little, Brown & Company, 1980, 436pp., index

Book Review

by

Norman Ansley

On January 22 and 23, 1976 the Masters & Johnson Institute held a conference and a transcription was published by Little, Brown & Company as *Ethical Issues in Sex Therapy and Research*. That meeting identified and discussed pertinent ethical issues. On January 25 to 27, 1978 a second meeting was held, and this Volume 2 is a transcript of the reports and discussions which led to the guidelines, developed by a task force in March 1978. A few readers will want to read part or all of the proceedings, but for most the value of the work is in the guidelines, pp. 406-420. The Guidelines include "Competence and Integrity of Sex Therapists;" "Confidentiality in Sex Therapy;" "Welfare of the Client;" "Welfare of Students and Trainees;" and "Welfare of the Research Subject." The guidelines are precise statements, not platitudes. For example, in describing their education, sex therapists and counselors should cite educational degrees "only when they have been received from an accredited agency or association." Although polygraph examiners are not mentioned, the guidelines include their activity when they are part of a sexual offender therapy team. Most of the guidelines presented in this Masters and Johnson book, with little rewriting, could be an appendix to the several guidelines for clinical polygraph examinations of sex offenders. The American Polygraph Association might have the Ethics and Grievance Committee rewrite the ethical guidelines for possible adoption, or the guidelines might be adopted by the organizations of examiners who test sexual offenders for full disclosure as an element of the required therapy.

* * * * * *

Lie Test

by

Leonard Harrelson with Nancy & Josh Gerow

Jonas Publishing, Fort Wayne, IN, 1998

Book Review

by

Robert Peters

For the past forty-five years, as director of the Keeler Polygraph Institute (KPI), Leonard Harrelson has been one of the most well known polygraph examiners in the world. His book, co-authored with Nancy & Josh Gerow is a terrific account of the many significant investigations and fascinating cases Len Harrelson has helped resolve. *Lie Test* is not only a history of Harrelson's experiences, but an important record of the critical role polygraph testing plays in the investigative process throughout the United States. For every polygrapher, *Lie Test* is a reminder that the profession they have chosen can be a doorway to a fantastic variety of newsworthy, mysterious situations that entail every aspect of the human condition.

Anyone who lived in Chicago during the 1970s will long remember the mysterious theft of over $4,000,000 in cash weighing over one-half ton from the highly secure vault of an armored car company. Len Harrelson remembers that event better than most. He helped the FBI solve that major crime. For the past 30 years, F. Lee Bailey has been the preeminent criminal defense attorney in the United States. Len Harrelson was an integral partner in a number of Bailey's most famous cases. Perhaps no single incident of the Vietnam War divided the nation more than the Mi Lai massacre and the subsequent trials of the two officers at the scene, William Calley and Ernest Medina. *Lie Test* describes Harrelson's elicitation of a confession of perjury by one of the prosecution's prime witnesses in the case of Ernest Medina. That admission eventually led to Medina's acquittal.

Harrelson provides the reader two chapters on methods of interrogation. One chapter deals with Harrelson's theories of the psychological basis of confessions and the qualities of an expert interrogator. He explains why empathy for the subject is the single most important characteristic of a good interrogator. The chapter on "How to Use Interrogation Analogies" will provide everyone who conducts interrogations some useful ideas that can be utilized immediately.

Unfortunately, for those with a strong interest in the techniques and theory of polygraph testing, *Lie Test* does not provide precise or sophisticated information regarding actual testing theory and technique. One might have expected the book to present a spirited argument as to the benefits of the Relevant/Irrelevant (R/I) technique. After all, KPI is the only polygraph school which presents the R/I as the primary testing procedure. However, the only reason Harrelson provides for using the R/I test is "because it is the techniques that Keeler used and because it has worked for him" (Harrelson). Equally surprising is the author's claim that "certain physiological changes occur uncontrollably when a person lies". This reader interprets the comment to mean there is a signature physiological reaction that occurs when a person lies. Most sophisticated analysts do not consider the signature lie

reaction a feasible concept. Harrelson offers a fairly extensive commentary on test question formulation. Unfortunately, the discussion never gets beyond the level of how to do it. This is disappointing, since some of the concepts are considerably different from those of other experts in the polygraph profession. Another surprising revelation is the fact that Harrelson never looks at a subject during the pretest interview. This seems at odds with the establishment of rapport, but that point is not addressed. Harrelson's longevity as an active polygrapher is astonishing. He has conducted over 50,000 exams in 40 years. A discussion regarding how he maintained his enthusiasm for a testing career that rivals Cal Ripken's performance would have been fascinating. Of all the test procedures recommended, the most surprising is Harrelson's belief that the examiner should be able to determine the subject's truthfulness after recording one polygraph chart. If the examiner is uncertain after the first chart, Harrelson recommends that the examiner interrogate the subject.

Anyone interested in polygraph needs to read *Lie Test.*

* * * * * *

Internet Reference Guide For Investigators: An Investigator's Guide to Sources of Information on the Worldwide Web

by

George Turner

Book Review

by

Venecca G. Green

An investigator's function is to gather information. Two elements affect how successful the investigator will be. The first is locating sources of information, and the second is the reliability of those resources. The Internet is the single most utilized information resource currently available. It contains an enormous amount of information. However, not everything contained on the Internet is reliable. *The Internet Reference Guide for Investigators: An Investigator's Guide to Sources of Information on the Worldwide Web* by George Turner, represents an attempt to provide the investigator with a tool to access reliable information via the Internet.

Turner's guide contains over 300 categories and approximately 1500 direct links to information related to the investigation profession. The contents are arranged alphabetically by topic. Turner is to be commended for the time spent researching the Internet to compile this document. This guide can save the proficient Internet user time and effort. It enables the user to go directly to a listed website, and it eliminates wading through page after page of websites. Going directly to a website without conducting a general search has its disadvantages, however. Limiting the number of sources potentially limits resources. The user can only hope that the sites listed are the most relevant. However, they may not be. It is unknown how the sites were chosen for inclusion in Turner's guide. More relevant resources may have been missed.

For example, Turner only lists three entries under the heading Terrorist Organizations. The AOL NetFind search engine was employed to conduct a general search using the key phrase *terrorist organization*, and over 800,000 sites were found. Certainly not all contain information relevant to the professional investigator, however, it is possible in a pool this large that the three listed are not the only relevant sources. Moreover, at this writing one of the three sites listed by Turner is no longer accessible.

It is impossible for the author to have included every possible topic or link because the Internet is so vast. Nonetheless, this guide is an excellent starting point for the investigator. It is a compilation of areas commonly investigated and resources commonly utilized by the investigator. The font size and the format of the document make it easy to read. It contains minimal editorial errors. One obvious error, however, is the misordering of entries in the table of contents: *Adoption–United States* was listed before *Acronym* and *Abbreviation Servers.* Although this guide was not thoroughly reviewed with a proofreader's eye, there are no glaring typographical or grammatical errors.

This writer evaluated the merits of the guide in part by considering the ease in which addresses were located in the guide, and the level of success in bringing up the listed websites onto the computer screen. To check the accuracy of the information, 50 addresses were randomly selected out of approximately 1500. Of the 50 addresses selected, only six were invalid or incorrect; an acceptable success record. On the negative side, not all users possess the same level of computer and Internet knowledge. For instance, the author uses the abbreviation URL throughout the document. The *Universal Resource Locator* (URL) is the same as the Internet address. Some users may be novices in the use of the computer, and the Internet is intimidating to others. To improve its user friendliness, it would have been desirable to include the following; a short glossary of terms, an illustration picturing the features of the typical Internet software interface, and help to finding the URL in the interface. Some users are less familiar with the modern keyboard, and they might benefit from guidance on locating unique character keys used in URLs.

The overall review and evaluation are positive. The $43.00 suggested list price cost should be weighed against the purchaser's skill in using common search engines. These search engines are free, but one needs to be adroit in their use to produce a list as distilled as is found in the Internet Reference Guide for Investigators.

One last caveat: the websites on the Internet are ever changing. The lifespan of this or any other hardcopy reference is adversely affected by those changes.

* * * * * *