Andrew B. Dollins,[1] Ph.D.; Victor L. Cestaro,[1] Ph.D. and Donald J. Pettit,[2] Ed.D.

Efficacy of Repeated Psychophysiological Detection of Deception Testing

REFERENCE: Dollins AB, Cestaro VL, Pettit DJ. Efficacy of repeated psychophysiological detection of deception testing. J Forensic Sci 1998;43(5):1016–1023.

ABSTRACT: Physiological measures were recorded during repeated psychophysiological detection of deception (PDD) tests to determine if reaction levels change with test repetition. Two groups of 22 healthy male subjects completed six peak of tension PDD tests on each of two test days. A minimum between test day interval of six days was maintained. The treatment group was programmed to respond deceptively to one of seven test questions while the control group was programmed to respond truthfully to all questions. The respiration and galvanic skin resistance (GSR) line lengths, GSR peak response amplitude and latency, and cardiovascular inter-beat-interval (IBI) were calculated for each response. Analyses indicated that, except for GSR peak response latency, differential physiological reactivity during a PDD test did not change significantly over repeated tests or days; there was a decrease in average respiration line lengths at the initial test(s) of each day; and differential changes in average respiration line length, GSR peak latency, and cardiovascular IBI responses corresponded to deception. Power analyses were calculated to assist in result interpretation. It is suggested that PDD decision accuracy, concerning subject veracity, should not decrease during repeated testing.

KEYWORDS: forensic science, psychophysiological detection of deception, peak of tension, habituation, repeated measures, respiration, galvanic skin resistance, heart rate, statistical analysis

The United States Department of Defense, various law enforcement agencies, and officers of the court routinely use a psychophysiological detection of deception (PDD) examination (1–3), commonly known as a ‘‘polygraph’’ or ‘‘lie-detector’’ test, to determine an individual’s truthfulness concerning topics of interest (4;5 pp. 1–8, 29–43). The theory underlying PDD is that physiologic reactivity, in response to the presentation of a stimulus, varies with the personal relevance of the stimulus and, more so, with attempts to conceal that relevance. The typical PDD examination is designed to elicit physiologic reactions from the examinee in response to questions regarding topic(s) of interest. Variability in galvanic skin resistance (GSR), respiratory rate and/or volume, and heart rate/blood pressure are typically assessed (visually) by PDD examiners in the field. An increase in reactivity (defined as a change in response rate and/or amplitude) in response to specific questions, is interpreted as indicative of the examinee’s truthfulness regarding the questions of interest.

Numerous valid criticisms have been expressed regarding the PDD process and associated assumptions (4–6). Among those are the criticisms of validity and reliability of results. Validity is defined (7, p. 749) as the degree to which a test measures what it is supposed to measure. Validity of a PDD examination would be measured as the degree of agreement between examiner decisions and ground truth (facts). Virtually all PDD studies attempt to assess the validity of PDD by comparing examiner decisions to ground truth. Definitions of ground truth range from experimental ‘‘programming,’’ [i.e., asking subjects to participate in mock crimes so guilt and innocence are known quantities; (8)] to decisions made by panels of experts who have reviewed case reports (9). While questions of validity are very important, they are moot if the data decisions are based on are not reliable. Reliability is defined (7, p. 629), as the degree to which a test measures the same thing consistently. As Lykken (4, p. 70), points out ‘‘A test can be highly reliable but have low validity; on the other hand, a test with low reliability cannot have high validity.’’ A test of PDD examination reliability would require testing the same individual on several occasions, using the same procedures. If the relationship between PDD examinee responses to different types of questions is not consistent among and between repetitions and different measures, it is unlikely that questions of validity can ever be properly addressed. There have been numerous studies of interexaminer reliability in evaluating physiological data collected during a PDD examination [e.g., (10–13)]. Such studies are important in that they examine the consistency of data interpretation among examiners. These studies have not, however, investigated the reliability of physiologic responses.

Few studies report results concerning the consistency of examinee responses. Results of an early exploratory study (14) designed to examine the GSR responses of ten male subjects using a variation of what is now labeled a stimulation card test (15, pp. 120–122), indicated that ‘‘one repetition of the detection procedure does not noticeably affect the success of the GSR as an indicator’’ of deception (14, p. 7). Results of a second similar study confirm this hypothesis ‘‘unless the subject is told that the first attempt was successful’’ (16, p. 11). A later study (17) employing a similar stimulation paradigm and GSR measure reported that identification of deception was improved by repeating the same question sequence ten times. Balloun and Holmes (18) recorded the responses of 16 male subjects during two five-question PDD examinations, separated by 30 seconds, administered using the Guilty Knowledge Questioning Technique (19). They found that responses were attenuated during the second administration of the

[1] Research psychologists, Department of Defense Polygraph Institute, Fort McClellan, AL.

[2] Program manager, Office of Criminal Investigations, The Food and Drug Administration, Rockville, MD.

Received 6 Aug. 1997; and in revised form 15 Dec. 1997; accepted 17 Dec. 1997.

test and suggest that repeated examinations may be invalid. Grimsley and Yankee (20) employed the Relevant/Irrelevant Question Technique to examine 80 male and female subjects on three occasions (separated by 24 hours). They found a nonsignificant decrease in accuracy between examinations 1 and 2, but no difference in accuracy between examinations 1 and 3. They concluded that overall accuracy rates are increased by evaluating multiple examinations. Yankee (21) used the Control Question Technique (22) and a somewhat more realistic paradigm to investigate the accuracy of repeated examinations. Subjects ($N = 72$) were examined on two occasions, separated by 24 hours. Half of the subjects were programmed ''guilty'' via participation in a mock crime. Yankee also reported a decline in accuracy between the two examinations, though smaller in magnitude than that reported by Balloun and Holmes (18).

Prior investigations of repeated PDD examinations have relied on visual inspection to evaluate physiological data and did not address the question of whether objectively measured response level differences occur during repeated examinations. The effect of a longer than 24-hour delay between successive PDD examinations has also not been examined. The current study was designed to examine relative levels of physiologic reactivity during repeated PDD examinations separated by more than six days. A relatively simple variation of the peak of tension paradigm was chosen under the assumption that the results would generalize to more complex paradigms which use questions of greater personal relevance.

Methods

Subjects—Forty-four, native English speaking, healthy males [mean age (standard deviation) = 29.2 (7.8) years; range = 19 to 47] participated in this study. They were military personnel or Department of the Army civilian employees and were not paid for their participation. Thirty-nine of the subjects had never participated in a PDD examination before. The remaining five had not participated in a PDD examination within the past two years. Thirty-five of the subjects reported themselves to be medication free. The remainder had ingested pain/relaxant (3 subjects), anti-inflammatory (1 subject), antibiotic (2 subjects), and antihistamine (3 subjects) medication within the 12-hour period prior to the examination.

Examiner—All PDD examinations were conducted by the same examiner, who had been trained at the United States Army Polygraph School and was certified by the United States Army to administer PDD examinations. He had administered approximately 500 field examinations during the five years prior to the study and was an instructor at the Department of Defense Polygraph Institute at the time of the study. The examiner was not aware of whether subjects belonged to the control or treatment groups.

Apparatus—Data were collected using a Lafayette (Lafayette, IN) Factfinder (Model no. 76740/76741) polygraph equipped with three multifunction Cardio | Aux | Pneumo | GSR modules (Model 76477-G), one GSR module (Model 76480-G), and one electronic stimulus marker module (Model 76351-GET). A circuit was added to the electronic stimulus marker module to allow control of the marker via signals from a computer RS-232 serial port. Lafayette sensors were used to measure GSR (Model 7664), respiration (Model 76513-1G and 76513-2B), and cardiovascular activity (Model 76530). Cardiovascular activity was recorded with the multifunction module selector set to Cardio-1.

An electronic circuit was designed and built in-house to amplify voltages from the Lafayette modules used to measure GSR, respiration, and cardiovascular activity. The amplification circuit contained potentiometers which could be used to adjust the preamplifier voltage offset. Amplification gains during testing were set at ×47 for the respiration channels, ×10 for the cardiovascular channel, and ×5 for the GSR channel. The amplified physiological signals were digitized using a Keithley Metrabyte (Taunton, MA) DAS-16F analog-to-digital converter installed in an IBM PS/Value Point (Armonk, NY) Model 433DX microcomputer. Software was written in-house to digitize the physiologic signals at a rate of 256 samples/s. A second micro-computer (Model 248, Zenith Data Systems, Chicago, IL), was used for question presentation to ensure that each question was presented with the same inflection, and at the same volume, each time it was repeated. The questions used throughout testing were digitized and recorded to computer hard disk using a Sound Blaster board (Model 16ASP, Creative Labs Inc., Milpitas, CA). Computer software was written in-house to allow the examiner to present questions (and to digitize physiological data) by moving a cursor on the computer screen. A parallel port interface (Speech Thing, Covox Inc., Eugene, OR), connected to a Radio Shack (Fort Worth, TX) integrated stereo amplifier (Model SA-155) and two speakers (Model Minimus-77), was used to present the questions. Subjects' verbal responses were recorded and examined as possible indexes of deception, as reported elsewhere (23).

PDD testing was conducted in a carpeted, 3.50 × 3.66 m partially sound-attenuated room. Each examination was recorded on video tape and monitored through a two-way mirror for quality control purposes. Subjects were seated in an adjustable-arm subject chair (Model 76871, Lafayette, IN) during PDD testing. The chair was positioned beside and slightly in front of the examiner's desk (Lafayette Model 76183). This position allowed the examiner to monitor the examinee's movements but not vice versa. The question presentation and data acquisition computers were positioned out of the examinee's sight during testing. The speakers, through which the questions were played, were located six feet behind, and one foot above, the back of the examinee's chair. The examinee's field of view, throughout testing, contained a wall of uniform color, a stationary video camera, and a piece of paper with numbers and words written on it (positioned above the video camera).

Procedure—Subjects were randomly assigned to the treatment or control groups, with the constraint that no more than three control or treatment group participants were tested consecutively. Twenty-two subjects were assigned to each group. Each subject participated in two examination sessions which were separated by at least six working days. Subjects completed six peak of tension PDD tests during each examination session.

Upon arrival at the test site, subjects were escorted to a secluded briefing room and asked to read a brief description of the research project. Subjects who indicated that they would participate were asked to read and sign a volunteer agreement affidavit. Their questions were then answered. A brief biographical/medical questionnaire was then completed, to ensure that the subjects were in good health and not currently taking medication which could interfere with the PDD examination results. Subjects then completed a number search task, which was referred to as an anagram task. During this task, the subject circled six sequences of a two-digit number which was repeated five consecutive times (in any direction) in a 20 × 30 matrix of two-digit numbers. The matrix consisted of numbers between 60 and 69 for the programmed deceptive subjects—who circled the number 64, and 80 to 89 for the programmed nondeceptive subjects—who circled the number 84. When the anagram task was completed, the subject was asked to write his name and the number he circled on two 7.62 × 12.70 cm cards. One card was retained by an investigator and the second concealed in the subject's pocket. The PDD examination procedure was briefly explained to the subject. It was emphasized that the subject should not reveal which number he had circled when completing the anagram task. It was further emphasized that he should remain relaxed, even if he felt himself begin to react (increased heart rate, perspiration on hands, tightening of occlusive cuff) during the examination. The subject was then escorted to the examination room and introduced to the examiner.

The examiner greeted each subject, then reviewed a biographical/medical questionnaire with him to ensure its accuracy. No other pretest questions were asked by the examiner. The examiner then briefly explained the sensors, procedures, and theory of PDD. It was carefully explained that the polygraph measured physiological reactions—and not deception per se. It was further explained that the subject's physiological responses were likely to change during deception. It was suggested that fear of detection during deception altered the normal physiological response pattern and that these changes may be evident in the signals recorded during the PDD examination. The examiner described this response as being similar to the fight-or-flight reaction used to describe a fear response during military training. The examiner then reviewed the questions to be asked during data collection by playing the recorded questions.

All questions asked by the subject were then answered. He was then seated in the examination chair and the sensors were attached. Respiration was monitored from the thoracic and abdominal areas. GSR was measured, without electrode paste, from the volar surface of the distal phalanges of the examinee's right hand index and ring fingers. Cardiovascular activity was monitored using an occlusive cuff placed around the upper left arm. The pneumo tube vents were closed and the amplifier DC offsets for the pneumo and GSR were adjusted to zero. The sensitivity of these recording channels was then adjusted on the polygraph. Next, the occlusive cuff was inflated to 90 mm Hg, massaged to remove wrinkles, then deflated to 48 mm Hg. The pressure was then adjusted, as necessary, to achieve a 2 mm Hg dial deflection between diastole and systole on the sphygmomanometer. The amplifier DC offset was then adjusted to zero, and polygraph sensitivity adjustments were made.

Each PDD test was composed of the following series of statements and questions:

X The test is about to begin.
01 Did you complete an anagram for the number 60?
02 Did you complete an anagram for the number 61?
03 Did you complete an anagram for the number 62?
04 Did you complete an anagram for the number 63?
05 Did you complete an anagram for the number 64?
06 Did you complete an anagram for the number 65?
07 Did you complete an anagram for the number 66?
XX The test is now complete, please continue to sit still while I turn the instrument off.

If the examiner judged the physiological signals recorded on the polygraph chart to contain artifacts, the previous question was repeated. The examiner played the prerecorded message ''please remain still'' if he judged that the examinee was producing unnecessary or excessive movements. When data collection for each test was completed, the pressure in the occlusive cuff was vented and the subject was instructed to ''please relax while I prepare for the next test.'' If subjects appeared to be sleepy, they were also reminded of the importance of the study and encouraged to remain alert. The next PDD test was begun approximately three minutes later. The occlusive cuff was inflated as described above, and DC offsets for the GSR and cardiovascular activity amplifiers were adjusted prior to beginning the next test. This process was repeated until six tests were completed, after which the sensors were removed. The subjects were then asked to read and sign a debriefing form, reminded to return the following week, and escorted out of the building.

Subjects returning for a second test session were escorted to a briefing room. They were reminded not to reveal the number circled during the previous session and asked to conceal the second card, indicating the number circled, in a pocket. They were then escorted to the examination room. The examiner again reviewed the biographical/medical questionnaire from their previous session to ensure that no significant changes had occurred. Six additional PDD tests were completed, as described above. When the examination was completed, the subject was thanked for his cooperation, asked to read and sign a second debriefing form, and escorted out of the building.

Data Reduction—The upper and lower pneumograph, GSR, and cardiovascular responses to each question were sampled at a rate of 256 samples/s for 14 s. Data sampling was initiated by the stimulus marker indicating that playback of the recorded question had ended. The data for each channel were smoothed to remove noise inherent in the instrument and/or amplifier used. Smoothing was implemented by substituting the average of the 50 points before and after a data sample (i.e., a running average of 101 data points) for that sample. The first and last 50 data points of each epoch were then omitted from the epoch. This smoothing procedure was empirically determined to be the optimal solution to reducing noise in the recorded signal.

The data collected during day 1, test 3, questions 61 through 64 were lost, due to experimenter error, for five subjects (3 deceptive and 2 nondeceptive). All responses were reviewed for movement artifact contamination by three psychophysiologists who were blind to the treatment condition of the collected sample. Responses identified as containing movement artifacts by two or more reviewers were marked as missing data and omitted from further processing. All responses with amplitudes that exceeded the limits of the analog-to-digital converter were marked as missing data.

The following statistics were calculated for the remaining 13.6 second epochs. Line length of the upper and lower pneumograph tracings (Pn1-LnL and Pn2-LnL, respectively), a technique introduced by Timm (24–26), and GSR (GSR-LnL) data were calculated using a between point interval of 0.00390625 units (i.e., 1/256). GSR peak amplitude (GSR-Amp) was calculated as the peak amplitude minus (0.5 * (Trough 1 + Trough 2) amplitudes). Troughs and peaks were identified as the first point where the subsequent 200 samples were greater (trough) or less (peak) than that point. If a peak was not identified within the first seven seconds of data sampling, the peak amplitude values for the epoch were set to 0.000. Trough 1 was the first trough occurring prior to the peak or the first data sample if a peak but no trough was located. Trough 2 was the first trough identified after the peak. GSR peak latency (GSR-Ltc) was calculated, in seconds, relative to the first data point collected, for analysis where peaks were found. If a peak was not identified, then the peak latency was considered missing data. The average heart rate inter-beat-interval (CRD-IBI) for each epoch was calculated by determining the latency between the first and last R-wave peak found during the 13.6 second epoch and dividing by the total number of peaks found during the epoch, minus one.

The mean and standard deviation of responses recorded under each condition of the independent variables (group, day, test, and question) were calculated and only values within two standard deviations of the mean were retained for further analysis. (Note that data previously described as missing were omitted from this calculation.) All missing data were replaced by means from the appropriate condition combination. The proportion of missing data for each measure—by deceptive/nondeceptive group, respectively, was: Pn1-LnL, 0.07/0.07; Pn2-LnL, 0.07/0.09; GSR-LnL, 0.14/0.10; GSR-Amp, 0.15/0.12; GSR-Ltc, 0.25/0.20; and, CRD-IBI, 0.05/0.07.

It was observed that more than 50% of the GSR line length and amplitude data were missing for two subjects in each group and that more than 50% of the GSR peak latency data were missing for six subjects in each group. The data for these subjects were not analyzed for these measures.

Data Analysis—Statistical analyses were calculated using SYSTAT for DOS (Version 5.0) and Windows (Version 5.04; SYSTAT, Inc.; Evanston, IL). The criterion for statistical significance was set at 0.05 or less throughout the result section. The Pn1-LnL, Pn2-LnL, GSR-LnL, GSR-Amp, GSR-Ltc, and CRD-IBI response measures were initially analyzed using a 2(between-group) × 2(within-day) × 6(within-test) × 6(within-question) repeated measure analysis of variance (ANOVA). As mentioned above: 22 subjects per group were included in the Pn1-LnL, Pn2-LnL, and CRD-IBI analyses; 20 subjects per group were included in the GSR-LnL and GSR-Amp analyses; and 16 subjects per group were included in the GSR-Ltc analysis. A completely within subjects 2(day) × 6(test) × 6(question) repeated measure ANOVA was subsequently calculated, where appropriate, to resolve group main and interaction effects (27, pp. 383–384). The degrees of freedom used in calculating each mean square error term and F statistic were reduced by the proportion of missing data for that measure. F statistic probabilities of repeated measure effects with more than two levels were corrected for violations of sphericity assumptions using the Greenhouse-Geisser (27, p. 523), epsilon (ϵ). Orthogonal planned comparisons (28, pp. 172–215), were used to evaluate significant ($p < 0.05$) test and question main effects. The comparisons chosen to evaluate test effects were: (*a*) test 1 versus tests 2, 3, 4, 5, and 6; (*b*) test 2 versus tests 3, 4, 5, and 6; (*c*) test 3 versus 4, 5, and 6; (*d*) test 4 versus tests 5 and 6; and (*e*) test 5 versus test 6. Significant question effects were evaluated by comparing the measures recorded in response to questions concerning the numbers 62, 63, 64, 65, and 66 to those recorded in response to the remaining questions. For example, the responses following the question concerning the number 62 were compared to those concerning the numbers 61, 63, 64, 65, and 66.

The statistical power of each ANOVA F-test was calculated to assess the probability that the null hypothesis of no difference between the treatment means would be correctly rejected when the hypothesis was false (29). Effect sizes were calculated as described by Cohen (30, pp. 531-545), then converted to the noncentrality parameter, lambda, by multiplying the squared effect size by the number of observations in each analysis (30, p. 550). It was necessary to convert effect sizes to a noncentrality parameter and calculate power directly rather than use Cohen's (30) effect size because the tables underestimate the power of factorial designs (31). The denominator degrees of freedom used in the power calculations were reduced by the percent of missing data, as described above. The power of each main effect and interaction was calculated using Laubscher's (32, Formula 6) square root approximation of noncentral F (26, p. 550). The results of this approximation were cross-checked with Bavry's (33) direct calculation of the noncentral F distribution.

The power of the 2 × 2 × 6 × 6 ANOVA day × test, group × day × test, day × question, group × day × question, test × question, group × test × question, day × test × question, and group × day × test × question F-tests to detect an effect size of 0.20 was at least 0.80-using a significance criterion of 0.05. The 2 × 2 × 6 × 6 ANOVA test, group × test, question, and group × question F-tests had a power of 0.80 to detect an effect size of 0.30 using a significance criterion of 0.05. The 2 × 2 × 6 × 6 ANOVA had relatively low power to detect group, test, and group × test effect sizes due to the small number of observations in these analyses. The power of reported statistical differences was at least 0.80 at a critical significance level of 0.05 or less. The degrees of freedom used during power calculation were adjusted to compensate for possible violation of sphericity assumptions using ϵ[(28), pp. 523, (34,35)] as suggested by Keppel (27, pp. 355–356).

Results

Pn1-LnL (Upper Pneumograph Line Length)—Pn1-LnL changed significantly over repeated tests [$F(5, 195) = 3.35$, $\epsilon = 0.70$]. Planned comparison results indicated that the Pn1-LnL measured during test 1 was longer [$F(1, 39) = 9.981$] than the average of those measured during tests 2, 3, 4, 5, and 6. This difference is illustrated in Fig. 1*a*. The group × question interaction was also significant [$F(5, 195) = 2.84$, $\epsilon = 0.60$].

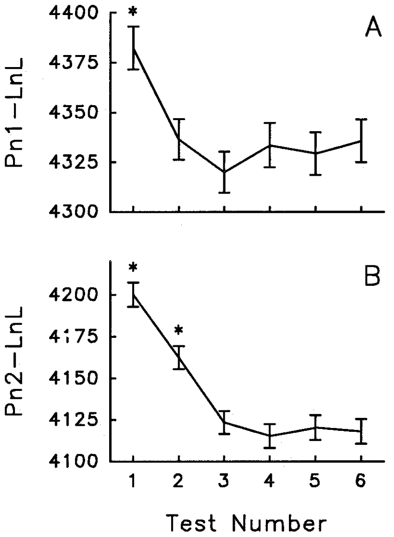

FIG. 1—*Pn1-LnL (*A*) and Pn2-LnL (*B*) responses averaged over the question, day, and group conditions. Vertical error bars represent standard error of the mean. Values marked with an asterisk (*) are significantly greater than subsequent values.*

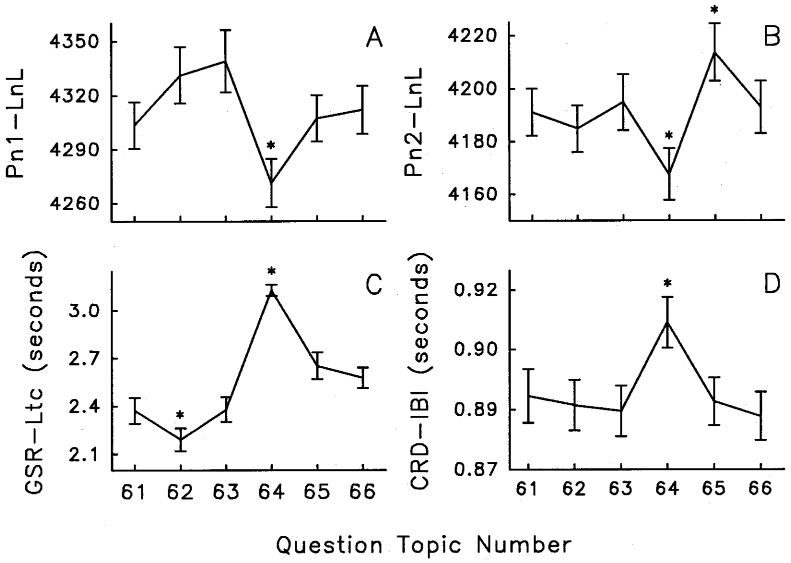

FIG. 2—*Deceptive subject Pn1-LnL (*A*), Pn2-LnL (*B*), GSR-Ltc (*C*), and CRD-IBI (*D*) responses averaged over tests and days. Vertical error bars represent standard error of the mean. Values marked with an asterisk (*) are significantly greater or less than the average of the remaining values.*

The deceptive and nondeceptive subject responses were analyzed separately to facilitate interpretation of the group × question interaction. A significant question effect [F(5, 98) = 3.59, ϵ = 0.39] was found among the deceptive subject responses, but not among those of the nondeceptive subjects. The results of subsequent comparisons among deceptive subject responses to questions, illustrated in Fig. 2a, indicated that the response to the question concerning the number 64 was shorter [F(1, 20) = 17.13] than the average of the remaining question responses.

Pn2-LnL (Lower Pneumograph Line Length)—Pn2-LnL responses measured from the deceptive subjects were longer than those measured from nondeceptive subjects [F(1, 39) = 9.40]. Pn2-LnL also changed significantly over repeated tests [F(5, 193) = 14.89, ϵ = 0.83]. Results of planned comparisons indicated that the Pn2-LnL measured during test 1 was longer [F(1, 39) = 46.03] than the average Pn2-LnL of subsequent tests, and that the Pn2-LnL measured during test 2 was longer [F(1, 39) = 18.02] than the average measured during tests 3, 4, 5, and 6, as illustrated in Fig. 1*b*. While a significant question effect was found [F(5, 193) = 3.76, ϵ = 0.82], the planned contrasts were all non-significant. The group × question interaction was significant [F(5, 193) = 5.07, ϵ = 0.82].

The deceptive and nondeceptive subject responses were analyzed separately to facilitate interpretation of the group × question interaction. A result of these analyses was that responses were shown to change significantly over repeated tests for both groups. The results of subsequent comparisons among tests showed the same pattern of significant effects as the overall analysis. Responses measured from the deceptive subjects differed significantly during question repetition [F(5, 97) = 5.52, ϵ = 0.66], while those measured from the nondeceptive subjects did not. Comparison results, illustrated in Fig. 2*b*, indicate that the deceptive subjects' Pn2-LnL response to the question concerning the number 64 was shorter [F(1, 19) = 9.05] than those in response to the remaining questions. In addition, the deceptive subjects' average Pn2-LnL response to the question concerning the number 65 was longer [F(1, 19) = 11.04] than the average Pn2-LnL responses to the remaining questions. Responses measured from nondeceptive subjects were also found to differ significantly during question repetition [F(5, 95) = 3.09, ϵ = 0.65], but no significant differences were found among the subsequent planned comparisons.

GSR-LnL (Galvanic Skin Resistance Line Length)—A significant group × day × chart interaction [F(5, 167) = 3.49, ϵ = 0.86] was found among the GSR-LnL measures, but simple effect analysis did not reveal where the differences occurred.

GSR-Amp (Galvanic Skin Resistance Amplitude)—The GSR-Amp measured from the deceptive subjects was greater [F(1, 33) = 10.35] than that measured from the nondeceptive subjects. GSR-Amp responses also changed significantly [F(5, 165) = 3.21, ϵ = 0.85] among repeated tests. Planned comparisons, however, failed to reveal any significant differences. Significant group × question [F(5, 165) = 13.29, ϵ = 0.79] and group × day × chart [F(5, 165) = 3.49, ϵ = 0.84] interactions were also found.

Separate analyses of the deceptive and nondeceptive subject GSR-Amp responses were calculated to facilitate interpretation of the group × question and group × day × chart interactions. A significant difference was found among the question responses of the nondeceptive subjects [F(3, 83) = 9.71, ϵ = 0.50]. Planned comparisons indicated that the GSR-Amp recorded in response to the question concerning the number 62 was greater than the average GSR-Amp response to the remaining questions [F(1, 16) = 11.34]. The GSR-Amp recorded in response to the question concerning the number 63 was less than the average response to the remaining questions [F(1, 16) = 13.51]. Significant differences were also found among the question [F(5, 80) = 6.92, ϵ = 0.74] and test [F(5, 80) = 2.81, ϵ = 0.74] responses of the deceptive subjects. The deceptive subject GSR-Amp response to the question concerning the number: 62 was smaller [F(1, 16) = 22.25] than that to the remaining questions; and, 66 was smaller [F(1, 16) = 16.79] than that to the remaining questions. No significant differences were found among the planned comparisons for the tests.

GSR-Ltc (Galvanic Skin Response-Response Latency)—A significant GSR-Ltc measure difference was found among responses to the questions asked during testing [$F(5, 115) = 9.29$, $\epsilon = 0.84$]. Comparisons indicate that response latencies to the question concerning the number 63 were shorter [$F(1, 23) = 11.93$] than those to the remaining questions. Response latencies to the question concerning the number 64 were longer [$F(1, 23) = 49.33$] than the average of those recorded in response to questions concerning the numbers 61, 62, 63, 65, and 66. The 2 × 2 × 6 × 6 ANOVA also indicated that there was a significant group × question effect [$F(5, 115) = 8.62$, $\epsilon = 0.84$].

Data recorded from the deceptive and nondeceptive groups were analyzed separately to assist in interpreting the significant group × question effect. Significant question effects were found for both the nondeceptive [$F(5, 60) = 5.01$, $\epsilon = 0.65$] and deceptive [$F(5, 56) = 19.69$, $\epsilon = 0.76$] subject responses. No significant differences were found among the question effect planned comparisons for the nondeceptive group. The deceptive subject GSR-Ltc response to the question concerning the number 62 was shorter [$F(1, 11) = 33.75$] than the average response to the remaining questions. The deceptive subject GSR-Ltc response to the question concerning the number 64 was longer [$F(1, 11) = 105.44$] than the average response to the remaining questions. These differences are illustrated in Fig. 2c.

A significant day × test × question effect [$F(25, 281) = 2.88$, $\epsilon = 0.35$] was found among the responses of the deceptive subjects. Separate analyses were calculated for the deceptive subject responses recorded during test days 1 and 2 to assist in interpreting this effect. These analyses indicated significant differences among the GSR-Ltc question responses for both day 1 [$F(5, 56) = 6.07$, $\epsilon = 0.71$] and day 2 [$F(5, 56) = 10.20$, $\epsilon = 0.68$]. Planned comparisons indicated that the deceptive subject GSR-Ltc responses to the question concerning the number 64, during day 1, were longer [$F(1, 11) = 41.77$] than the average response latency to the remaining questions. Comparisons for deceptive responses measured during day 2 indicate that responses to the question 62 were shorter [$F(1, 11) = 59.36$] than the average response latency to the remaining questions and that responses to the question concerning the number 64 were longer [$F(1, 11) = 41.37$] than the average response latency to the remaining questions.

A significant test × question effect was found among the deceptive subjects GSR-Ltc responses during test day 2 [$F(25, 281) = 2.22$, $\epsilon = 0.32$]. Each test was analyzed separately to assist in interpreting this difference. No significant differences were found among the question responses recorded during tests 1 and 5. The analyses indicated that there were significant differences among responses recorded to questions during tests 2, 3, 4, and 6. Contrasts indicate that the GSR-Ltc responses to the question concerning the number 64 were significantly longer than the average of those recorded in response to the remaining questions during tests 2, 3, 4, and 6. GSR-Ltc responses to questions concerning the numbers 62 and 66 recorded during test 4 and to questions concerning the number 62 recorded during test 6 were significantly shorter than the average of the responses recorded during the remaining questions.

Separate analyses were calculated for the nondeceptive subject responses recorded during test days 1 and 2 to assist in interpreting a significant day × test effect result found during the analysis of nondeceptive subject GSR-Ltc responses [$F(5, 60) = 2.72$, $\epsilon = 0.68$]. No significant test, question, or test × question effects were found among the nondeceptive subject GSR-Ltc responses recorded during day 1. Nondeceptive subject responses on day 2 were, however, found to differ significantly among questions [$F(5, 60) = 4.46$, $\epsilon = 0.623$]. Planned comparisons indicate that the GSR-Ltc response latency to the question concerning the number 63 was shorter than the average response latency to the remaining questions [$F(1, 12) = 13.71$].

CRD-IBI (Cardio Channel Average Inter-beat-interval)—A significant CRD-IBI measure difference was found among responses to the questions asked during testing [$F(5, 197) = 4.27$, $\epsilon = 0.53$]. Comparisons indicate that the CRD-IBI measured in response to the question concerning the number 64 was longer [$F(1, 39) = 14.80$] than those to the remaining questions. The analysis also indicated significant group × question [$F(5, 197) = 3.41$, $\epsilon = 0.53$], group × day × test [$F(5, 197) = 3.06$, $\epsilon = 0.83$], and group × test × question interactions [$F(25, 987) = 1.93$, $\epsilon = 0.45$].

Separate analysis of the data recorded from the deceptive and nondeceptive subjects indicated no significant differences among the nondeceptive subject responses as a function of the independent variables manipulated. A significant question effect was, however, found among the deceptive subject responses [$F(5, 99) = 5.84$, $\epsilon = 0.54$]. Planned comparisons indicated that the deceptive subjects CRD-IBI response to the question concerning the number 64 was longer than the average response CRD-IBI to the remaining questions, as illustrated in Fig. 2d.

Discussion

These results suggest that during repeated administration of PDD tests: there is a consistent change in average Pn1-LnL and Pn2-LnL; differential Pn1-LnL, Pn2-LnL, and CRD-IBI reactivity during a PDD test does not change during repeated tests or days; and, average physiological reactivity of deceptive subjects changes during deception while that of nondeceptive subjects does not. When interpreting these results it is important to remember that the power of each significant statistical effect was 0.80 or greater and that the power of the non-significant statistical tests to detect an effect of size 0.30 at the 0.05 significance level was also 0.80 or greater (with exceptions noted above). The power analysis provides the probability (0.80 or greater) that the null hypothesis is correctly rejected when a significant effect was observed, as well as the probability (0.80 or greater) than an effect size of 0.30 would have been correctly detected.

Perhaps the most interesting result of this research is not the significant results which were obtained, but those that were not. All day × test, day × question, test × question, and day × test × question interactions were non-significant. This suggests that the pattern and/or variability of measured physiologic responses to the questions asked during each PDD test did not change significantly during repeated administration of the tests, nor did the response pattern change significantly between days 1 and 2, with the exception of GSR-Ltc responses. While subject veracity was not directly examined, this result is interpreted as supporting earlier reports (14,17,20,21) that there were no statistically significant differences in the detection of veracity with repeated testing. While veracity detection rates were not determined, the conclusion that differential responding does not change with question series repetition supports the proposal that decision consistency should not change with repeated testing (17,20,36).

The results of some investigations into the effect of repeated question series administration on skin resistance and/or conductance responsivity do not support those of this study (16,18,36,37) while those of others do (38,39). This is a difficult issue to resolve due to methodological differences in the: response requirements; question repetition patterns and procedures; and, data reduction, evaluation, and analysis techniques. It is also possible that the response strengths measured during this study decreased with repetition, but the decrease was too small to be statistically detected. It is likely, however, that such small changes would be of little interest. Further research should be conducted to address these issues. Average Pn1-LnL and Pn2-LnL response levels measured during the first test, averaged over groups, days, and questions, were found to be significantly greater than the average of the subsequent tests, as illustrated in Fig. 1. No statistically significant difference was found between Pn1-LnL measures recorded during tests 2 through 5 and the average of subsequent tests. The Pn2-LnL measure recorded during test 2 was significantly greater than those recorded during tests 3 through 6, but measures recorded during tests 3 through 5 were no different from those recorded during subsequent tests. A similar shift in skin conductance following repeated testing has been reported elsewhere (36). The decrease in average response levels observed during the initial stages of repeated testing, in the absence of within test response attenuation, may be a variation of the phenomenon of differential autonomic responsivity (37).

Results of the data analyses indicate that there were no statistically significant main or interaction effects related to the questions asked among the average nondeceptive subject Pn1-LnL, Pn2-LnL, and CRD-IBI responses. The average deceptive subjects' deceptive responses were shorter in Pn1-LnL and Pn2-LnL, longer in GSR-Ltc, and longer in CRD-IBI than the average of their nondeceptive responses. These results confirm that, on the average, a pattern of differential responding occurs during deception that does not occur when deception is not present. While pneumo line lengths and heart rate are not normally evaluated when scoring PDD examinations, perhaps polygraphs used for PDD should be modified to display this information.

While significant differences were found among the deceptive subjects' GSR-Amp responses to the questions asked, the deceptive response was not significantly different from the average nondeceptive response. This is surprising when one considers results of studies reporting high-veracity detection accuracy rates based exclusively on electrodermal activity scores (40–43). However, close examination of these reports suggests that differences in methodology and evaluation techniques could account for the differences between the current results and earlier reports. While a field polygraph was used in the current study, the operator sensitivity adjustments were bypassed. Skin resistance changes were amplified by a fixed-gain linear amplifier adjusted to remain within the range limits of an analog-to-digital converter, which did not compensate for changes in tonic skin resistance, possibly contributing to the failure to find significant differences among GSR-Amp measures during deception, in this study.

It should, however, be noted that 9% and 27% of the subjects were dropped from the GSR-Amp and GSR-Ltc analyses, respectively, due to insufficient data caused, primarily, by failure to obtain quantifiable subject responses. The percentages of missing Pn1-LnL, Pn2-LnL, and CRD-IBI data, which were collected simultaneously with the GSR data, were not sufficiently large to necessitate removal of subjects from the analyses. This observation is interpreted as suggesting that the exclusive or disproportionately high reliance on GSR response scores when interpreting the results of PDD examinations may lead to excessive errors. This suggestion is not new, but simply reinforces the statement presented to the Committee on Government Operations over 20 years ago that ''most examiners agree that the galvanic skin response is the least accurate, and should be ignored when a conflict (among the three channels) occurs'' (44).

In summary, three conclusions are derived from the results of this research. First, a consistent change was observed in average Pn1-LnL and Pn2-LnL responses, but not the GSR-Amp, GSR-LnL, GSR-Ltc, and CRC-IBI responses as the test was repeated. This pattern did not change significantly between test days one and two. Second, the average physiological response variability measured during a PDD test did not change over repeated tests. Finally, the Pn1-LnL, Pn2-LnL, GSR-Ltc, and CRD-IBI responses of deceptive subjects, averaged over repeated test administrations, changed during the deceptive response, relative to nondeceptive responses. No such systematic changes were found among the responses of the nondeceptive subjects. These data are interpreted as suggesting that decision consistency should not be significantly affected by repeated (up to six) administrations of the question series during a PDD examination. This conclusion is supported by reports by others (17,20,36). We further suggest that changes in heart rate inter-beat-interval, measured using an occlusive cuff as described, and pneumo line length are reliable response measures which may be accurately interpreted as indicating deception.

Acknowledgments

The authors wish to express special thanks to: the volunteers who participated in the study; Charlene L. Stephens, Master Sergeant Randall S. Reynolds, United States Air Force, and the support staff at the Department of Defense Polygraph Institute (DoDPI), without whom this study could not have been completed. The authors would also like to thank Chief Warrant Officer Four Madison L. Mumbauer, United States Army, for his editorial assistance, and the Director of the DoDPI, Dr. William J. Yankee, for his continued support throughout the project. This study was supported by funds from the DoDPI as project DoDPI93-P-0025. The views, opinions, and findings contained in this paper are those of the authors, and should not be construed as official or reflecting official Department of Defense policy. Excerpts of this manuscript were presented at the Annual Meeting of the American Academy of Forensic Sciences, New York, NY, March 1997.

References

1. Krapohl D, Sturm S. Terminology reference for the science of psychophysiological detection of deception. Chattanooga (TN): American Polygraph Association, in press.
2. Podlesny J, Raskin D. Physiological measures and the detection of deception. Psychol Bull 1977;34:782–99.
3. Yankee WJ. The current status of research in forensic psychophysiology and its application in the psychophysiological detection of deception. J Forensic Sci 1995;40:63–68.
4. Lykken DT. A tremor in the blood: Uses and abuses of the lie detector. New York: McGraw-Hill, 1981;1–8.
5. Office of Technology Assessment. Scientific validity of polygraph testing: A research review and evaluation—a technical memorandum. Washington (DC): Office of Technology Assessment; 1983 Nov. Report No.: OTA-TM-H-15, 29–43.
6. Furedy JJ. Lie detection as a psychophysiological differentiation: some fine lines. Coles MGH, Donchin E, Porges SW, editors. Psychophysiology: Systems, Processes, and Applications. New York: Guilford Press, 1986;683–701.
7. Campbell RJ. Psychiatric Dictionary (6th ed). New York: Oxford University Press, 1989.
8. Barland GH, Raskin DC. An evaluation of field techniques in detection of deception. Psychophysiol 1975;12:321–30.

9. Bersh PJ. A validation study of polygraph examiner judgments. J Appl Psychol 1969;53:399–403.
10. Horvath F, Reid J. The reliability of polygraph examiner diagnosis of truth and deception. J Crim Law Crim & Pol Sci 1971;62: 276–81.
11. Horvath F. The effect of selected variables on interpretation of polygraph records. J Appl Psychol 1977;62:127–36.
12. Hunter FL, Ash P. The accuracy and consistency of polygraph examiners' diagnoses. J Pol Sci & Adm 1973;1:370–5.
13. Slowik SM, Buckley JP. Relative accuracy of polygraph examiner diagnosis of respiration, blood pressure, and GSR recordings. J Pol Sci & Adm 1975;3:305–9.
14. Ellson DG, Davis RC, Saltzman IJ, Burke CJ. Accuracy of detection and the effect of repetition. A report of research on detection of deception. Bloomington (IN): Indiana University, Department of Psychology; 1952 Tech Rep Contract Number N60NR-18011.
15. Abrams S. The complete polygraph handbook. Lexington (MA): Lexington Books; 1989.
16. Elaad E, Ben-Shakhar G. Effects of motivation and verbal response type on psychophysiological detection of information. Psychophysiol 1989;26:442–51.
17. Lieblich I, Naftali G, Shumueli J, Kugelmass S. Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. J Appl Psychol 1974;59:113–5.
18. Balloun KD, Holmes DS. Effects of repeated examinations on the ability to detect guilt with a polygraphic examination: A laboratory experiment with a real crime. J Appl Psychol 1979;64:316–22.
19. Lykken DT. The validity of the guilty knowledge technique: The effects of faking. J Appl Psychol 1960;44:258–62.
20. Grimsley DL, Yankee WJ. The effect of multiple retests on examiner decisions in applicant screening polygraph examinations. Charlotte (NC): A. Madley Corporation; 1986 National Security Agency Contract No. MDA 904–85-C-A962.
21. Yankee WJ. Test and re-test accuracy with a psychophysiological detection of deception test. Ft. McClellan (AL): Department of Defense Polygraph Institute; 1993. Adopted from Yankee WJ, Grimsly DL. The effect of a prior polygraph test on a subsequent polygraph test. Charlotte (NC): University of North Carolina; 1986 National Security Agency Contract No. MDA 904–84-C-4249.
22. Reid JE. A revised questioning technique in lie detection tests. J Crim Law & Crim 1947;37:542–7.
23. Cestaro VL, Dollins AB. An analysis of voice responses for the detection of deception. Polygraph 1996;25:15–34.
24. Timm H. The effect of placebos and feedback on the detection of deception (US Department of Justice Contract No. 78-NI-AX-0028). East Lansing (MI): Michigan State University; 1979.
25. Timm H. Analyzing deception from respiration patterns. J Pol Sci & Adm 1982;10:47–51.
26. Timm H. Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. J Appl Psychol 1982;67:391–400.
27. Keppel G. Design and analysis, a researcher's handbook (3rd ed.). Englewood Cliffs, (NJ): Prentice Hall, 1991.
28. Winer BJ. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.
29. Williams RH, Zimmerman DW. Statistical power analysis and reliability of measurement. J Gen Psychol 1989;116:359–69.
30. Cohen J. Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale (NJ): Lawrence Erlbaum Associates, 1988.
31. Koele P. Calculating power in analysis of variance. Psychol Bull 1982;92:513–6.
32. Laubscher NF. Normalizing the noncentral t and F distributions. Ann Math Statist 1960;31:1105–12.
33. Bavry JL. Stat-Power statistical design analysis system. Chicago: Scientific Software, Inc., 1991.
34. Geisser S, Greenhouse SW. An extension of Box's results on the use of the F distribution in multivariate analysis. Ann Math Statist 1958;29:885–91.
35. Greenhouse SW, Geisser S. On methods in the analysis of profile data. Psychometrika 1959;24:95–112.
36. Iacono WG, Boisvenu, GA, Fleming JA. Effects of diazepam and methylphenidate on the electrodermal detection of guilty knowledge. J Appl Psychol 1984;69:289–99.
37. Ben-Shakhar G, Lieblich I. The dichotomization theory for differential autonomic responsivity reconsidered. Psychophysiol 1982; 19:277–81.
38. Furedy JJ, Ben-Shakhar G. The roles of deception, intention to deceive, and motivation to avoid detection in the psychophysiological detection of guilty knowledge. Psychophysiol 1991;28:163–71.
39. Furedy JJ, Gigliotti F, Ben-Shakhar G. Electrodermal differentiation in deception: the effect of choice versus no choice of deceptive items. J Psychophysiol 1994;18:13–22.
40. Iacono WG, Cerri AM, Patrick CJ, Fleming JAE. Use of antianxiety drugs as countermeasures in the detection of guilty knowledge. J Appl Psychol 1992;77:60–4.
41. Kugelmass S, Lieblich I. The effects of realistic stress and procedural interference in experimental lie detections. J Appl Psychol 1966;50:211–6.
42. Podlesny J, Raskin D. Effectiveness of techniques and physiological measures in detection of deception. Psychophysiol 1978;15: 344–59.
43. Thackray R, Orne M. A comparison of physiological indices in detection of deception. Psychophysiol 1968;4:329–39.
44. Committee on Government Operations. The use of polygraphs and similar devices by federal agencies (Hearings before a Subcommittee of the Committee on Government Operations, House of Representatives, 93rd congress, 2nd session). Washington (DC): U.S. Government Printing Office, 1974.

Additional information and reprint requests:
Andrew B. Dollins, Ph.D.
Building 3195
Ft. McClellan, AL 36205