Effects of Prior Demonstrations of Polygraph Accuracy on Outcomes of Probable-Lie and Directed-lie Polygraph Tests

John C. Kircher[1], Ted Packard[1], Brian G. Bell[1] and Paul C. Bernhardt[1]

Abstract

The present study tested if the stimulation pretest improves the accuracy of probable-lie and directed-lie tests. 336 men and women were recruited from the general community and were paid $30 to participate in a mock crime experiment. Equal numbers of males and females were assigned to one of 16 cells in a 2 X 2 X 4 factorial design, with two levels of Guilt (guilty and innocent), two levels of Test Type (probable-lie and directed-lie), and four variants of pretest procedures. Half of the participants were guilty and half were innocent of committing a mock theft of $20 from a purse, and all participants were promised and paid a $50 bonus if they could convince the polygraph examiner that their innocence. Half of the participants were given probable-lie tests and half were given directed-lie tests. 120 participants were not given the stimulation pretest (no-pretest). 120 participants were given the stimulation test and told that the polygraph clearly revealed their deception (effective-feedback). 48 participants were given the stimulation test but no feedback about the outcome (no-feedback). The remaining 48 participants were given the pretest and told that the polygraph failed to reveal their deception (ineffective-feedback).

As compared to the no-pretest control condition, the combination of the pretest and effective feedback increased the accuracy of decisions from 77% to 90% for the probable-lie test and from 75% to 83% for the directed-lie test. Additional comparisons revealed that the observed improvement in decision accuracy for probable-lie tests was due to the pretest and not the feedback. However, for the directed-lie test, decision accuracy was higher when the pretest was followed with effective feedback than when it was not.

There were no significant differences between probable-lie and directed-lie tests in the accuracy of decisions by independent numerical evaluators or computer diagnoses, but respiration measures were more diagnostic for the probable-lie test. Among non-traditional physiological measures, skin potential was found to be as diagnostic as skin conductance, and measurements of blood pressure from a finger were found to be at least as diagnostic as the cardiograph.

Introduction

The present study had three major objectives. The primary objective was to assess the effects of the stimulation test on the accuracy of subsequent polygraph examinations. The second objective was to assess the relative effectiveness of probable-lie (PL) and directed-lie (DL) comparison question tests. The third objective was to assess the ability of several new physiological measures to discriminate between truthful and deceptive individuals.

Background

Field polygraph examiners commonly administer a stimulation test prior to conducting specific-incident polygraph examinations. Although there are several variants of the stimulation test (Abrams, 1989), the numbers test is the approach most often borrowed from the field for use in laboratory research on polygraph techniques (Raskin, 1989). The

Editor's note: This research was funded by the DoD Polygraph Institute (now the Defense Academy for Credibility Assessment), and the final report was received on October 15, 2001. The original report can be retrieved from the Defense Technical Information Center. The report identification is DTIC AD Number A404128

[1]University of Utah, Salt Lake City, Utah 84112

polygraph examiner begins by telling the subject that each individual is unique and that each individual shows a different pattern of physiological response when they lie during a polygraph test. The subject is then told that a preliminary test with numbers will be conducted that will allow the examiner to see what the subject's physiological responses look like when they lie and when they tell the truth. The subject chooses a number between 3 and 6 and is instructed to deny having selected that number when asked during the test. The subject is then asked about the numbers 1 through 7 while their physiological responses are recorded on the polygraph. Subjects are deceptive when they answer "No" to the question about the selected number and are truthful when they answer "No" to questions about the other numbers. At the end of the pretest, the subject is told, regardless of the actual result, that they showed a strong physiological reaction when they lied and no significant reactions when they were truthful. Consequently, the subject is told, there should be no problem in determining if they are truthful or deceptive on the polygraph test.

The purpose of the pretest is to lead the individual to believe that the polygraph will reveal when they are truthful and deceptive. If a deceptive individual believes that the polygraph will reveal that their statements about the matter under investigation are false, then they should be concerned or fearful when asked about the crime and should react strongly to those questions. On the other hand, truthful subjects may be concerned about the possibility of failing the test because the polygraph might not indicate that their answers are truthful. A demonstration that the polygraph accurately distinguishes between truthful and deceptive responses should help to allay those concerns. In general, it is assumed that the pretest affects subjects' beliefs about the accuracy of the polygraph technique, and the effect is to facilitate accurate discrimination between truthful and deceptive answers to questions on the test.

Results of research on the effects of the stimulation test are mixed. Ellson, Davis, Saltzman, and Burke (1952) examined the effects of demonstrated effectiveness on the accuracy of subsequent polygraph outcomes in a mock crime experiment. After one trial, some subjects were given feedback that their attempts to deceive were detected, and other subjects were told their lies were not detected. Feedback to subjects that their lies were detected on the first trial made it more difficult to detect deception on subsequent trials. Conversely, feedback to other subjects that their lies were not detected improved detection on subsequent trials. Davis (1961) suggested that the guilty subjects might become less physiologically reactive and less detectable if the feedback they receive convinces them that their lies are clearly revealed by the polygraph. These early findings, and Davis' conclusions, are contrary to the assumption that a pretest demonstration of polygraph accuracy enhances discrimination between truthful and deceptive responses.

Subsequent research by Bradley and Janisse (1981), however, supports the use of the stimulation test. Bradley and Janisse obtained measures of skin resistance, heart rate, and pupil diameter from 192 college students, half of whom had committed a mock theft. Prior to a probable-lie comparison question test, participants were given three card tests and feedback that they had been detected on none, one, two, or all three of the card tests. As predicted, subsequent discrimination between truthful and deceptive subjects on measures of skin resistance increased monotonically with increases in the level of demonstrated effectiveness, although there were no effects on heart rate or pupillary responses.

Of all the measures obtained in field polygraph examinations, electrodermal measures carry the greatest weight in the decision processes of expert polygraph examiners and computer models (Kircher & Raskin, 1981; 1988; Raskin et al., 1988). Therefore, the findings obtained by Bradley and Janisse (1981) for skin resistance support the use of a stimulation test to improve discrimination between truthful and deceptive individuals, and they are often cited in the scientific literature as a justification for the continued use of the stimulation test in field polygraph examinations (e.g., Horowitz, Kircher, Honts, & Raskin, 1997; Podlesny & Truslow, 1993; Raskin, 1989; Saxe, Dougherty & Cross, 1985).

Despite both empirical and theoretical justifications for the use of the stimulation test, several questions may be raised about the relevance of the results from the Bradley and Janisse (1981) study for field polygraphy. Bradley and Janisse recruited only white, male college students for their experiment. It is unknown if the observed effects of demonstrated effectiveness will generalize to a population of individuals who vary in age, sex, intelligence, and ethnicity.

The incentives that were offered by Bradley and Janisse to participants to pass their polygraph tests were unrealistic. Half of the students in the guilty and innocent treatment conditions were threatened with an electric shock if they failed their test. However, this manipulation had no effect on self-reports of state anxiety or electrodermal measures. Otherwise, guilty subjects were offered $1 to beat the test, and innocent subjects were given no incentive to pass the test.

Participants' involvement in the mock crime was also limited. The participants received written instructions in a room on a university campus. They were told to imagine that the room was part of a store. An envelope that contained $1 was placed on a shelf under the desk where they received their instructions. Subjects in the guilty condition were instructed to "steal" the money. In a meta-analysis of mock crime experiments (Kircher, Horowitz, & Raskin, 1988), the incentives and level of subject involvement achieved in the Bradley and Janisse study were rated as relatively weak and unrepresentative of field situations. At this point, we do not know if the effects reported by Bradley and Janisse would be obtained if the subjects, incentives, and level of personal involvement achieved in their experiment had more closely approximated field conditions.

The present study employed a mock crime scenario that differed in several respects from the one used by Bradley and Janisse. We borrowed the protocol from one of our previous experiments (Kircher & Raskin, 1988). We used this particular research paradigm because the data from the Kircher and Raskin experiment have been compared to verified truthful and deceptive charts recorded by U. S. Secret Service examiners in a subsequent field validity study (Raskin, Kircher, Honts, & Horowitz, 1988). Comparisons of covariance and mean structures for the laboratory and field data suggest that our mock crime procedures yield patterns of physiological reactivity that closely resemble those obtained in the field (Kircher, Raskin, Honts, & Horowitz, 1994).

The present study differed in another respect from the one conducted by Bradley and Janisse (1981). We included a condition in which subjects were given a preliminary numbers test, but they were not provided with feedback regarding the outcome (cf., Podlesny & Truslow, 1993). Inclusion of the no-feedback condition allowed us to test if the mere presentation of the numbers test affects detection rates independently of the feedback provided about the outcome. This aspect of the design allowed us to test if the numbers test improves subsequent detection rates, not because it affects subjects' beliefs, but because it provides an opportunity for the subject to habituate to the sensors, task, or setting.

We also explored the possibility that people who have a strong need for social approval, are more trusting, or are more anxious are more affected by demonstrations of polygraph effectiveness than people who are less compliant or anxious. If the stimulation test improves polygraph accuracy by altering examinees' beliefs, then individual differences in the need to appease others, interpersonal trust, or anxiety may moderate the effect of the stimulation effects on polygraph accuracy. To our knowledge, the need for social approval and interpersonal trust has not been examined in prior research with the polygraph. Previously, Honts, Raskin, and Kircher (1986) observed no relationship between anxiety and detectability, but they did not manipulate prior demonstrations of polygraph accuracy and, therefore, could not test if anxiety interacts with demonstrations of polygraph effectiveness.

Another objective of the present study was to assess the relative effectiveness of probable-lie and directed-lie comparison question tests. Federal polygraph examiners are currently trained at the Department of Defense Polygraph Institute (DODPI) to administer the traditional probable-lie test as well as the more recent directed-lie comparison test. Probable-lie questions are designed to capture the attention and cause concern in examinees who answer the relevant questions truthfully, and their effectiveness depends on a psychological manipulation of examinees by the polygraph examiner (Raskin, Kircher, Horowitz, & Honts, 1989). This manipulation is difficult to standardize and is ineffective with some truthful people who fail the test because they react more strongly to the relevant questions (false positive errors).

There is also a problem with deceptive individuals who show stronger reactions to the probable-lie questions even though they are lying to the relevant questions. This might occur if the comparison questions encompass transgressions of greater concern to these examinees than the accusations embodied in the relevant questions. Such problems may have contributed to the high rate of false negative errors obtained in a security screening study by Barland, Honts, and Barger (1989), since a surprising number of their deceptive examinees admitted to having committed actual security violations in the past. Because there are problems with probable-lie questions, use of directed-lie questions by federal polygraph examiners has increased in recent years (1997 DoDPI Seminar, Ft. McClellan, AL).

Fuse and Hitchcock developed directed-lie questions for use in Vietnam in 1966 (Gaschler, personal communication), and the first written description of the directed-lie technique appeared in 1982 (Fuse, 1982). The rationale for using directed-lie questions is similar to the rationale for using traditional probable-lie questions (Raskin et al., 1989). Although the directed-lie test is administered and scored in the same manner as a probable-lie test, directed-lie tests require less psychological manipulation of the examinee and less examiner skill and sophistication. Directed-lie questions have greater face validity, they are less personally intrusive and embarrassing to the examinee, and they are more easily standardized. Since repeated testing of the same person may pose problems for the use of traditional probable-lie control questions, directed-lie questions have additional advantages in the security screening context.

Directed-lie tests have many advantages, but relatively little is known about their validity. To date, only two studies of the directed-lie test have been published (Honts & Raskin, 1988; Horowitz, Kircher, Honts, & Raskin, 1997), and both studies suggest that the directed-lie test is at least as accurate as the probable-lie test. The present study compared the outcomes obtained with probable-lie questions to those obtained with directed-lie questions. In so doing, the present study provides additional data on the validity of the directed-lie test as well as an evaluation of the effects of the stimulation test on the outcomes of directed-lie tests.

A tertiary objective of this study was to explore several new physiological measures that may improve discrimination between truthful and deceptive individuals. In certain contexts, the accuracy of decisions by the computer and polygraph examiners may exceed 90% on both truthful and deceptive subjects (Raskin et al., 1989). However, decisions by computers and human experts are not perfect. The accuracy of computer diagnoses may be improved by devising new ways of recombining measurements obtained from the physiological waveforms recorded by standard field polygraph instruments (e.g., Honts, 1992). Field polygraphs record thoracic and abdominal respiration, skin resistance or conductance, cardiovascular activity, and finger pulse amplitude. Improved accuracy may also be achieved by identifying new sources of diagnostic information in physiological measures that are not currently recorded by field polygraph instruments.

We already know that computers can be used to extract a large number of highly diagnostic measures from each of the standard physiological waveforms (Kircher & Raskin, 1988; Podlesny & Raskin, 1978; Raskin et al., 1988). In one study, we found that 11 of 12 characteristics of skin conductance responses (amplitude, duration, frequency, etc.) reliably discriminated between truthful and deceptive subjects (Kircher & Raskin, 1988). However, multiple measurements obtained from a single response waveform, such as skin conductance, are highly inter-correlated. Consequently, a single well-chosen measure captures virtually all of the diagnostic information available in the set of component measures. On the other hand, relatively low correlations are found among measurements obtained from different channels of physiological activity (e.g., skin conductance and respiration). Because such measures are relatively independent, they provide complementary rather than redundant sources of diagnostic information. Weighted combinations of such measures are significantly more diagnostic than any one measure taken individually.

Several years ago, we completed a pilot study that was designed in part to evaluate the utility of several new physiological measures for detecting deception (Raskin & Kircher, 1990). Several interesting findings emerged. The single most diagnostic measure was derived from skin potential responses. Among standard measures, the amplitude of skin conductance responses is typically the most valid measure. The results obtained for skin potential are especially encouraging because different physiologic mechanisms are thought to underlie skin potential and skin conductance responses (Venables & Christie, 1973). This suggests that the two measures are likely to be somewhat independent.

Raskin and Kircher (1990) also derived measures of vagal tone from the electrocardiogram (EKG) for each of four 5-second epochs following stimulus (question) onset. Vagal tone is a measure of the inhibitory effects of the vagus on heart rate (Porges et al., 1980). During the fourth epoch, deceptive subjects showed less vagal tone in response to relevant questions than to probable-lie questions. In contrast, truthful subjects showed less vagal tone in response to probable-lie questions than to relevant questions. The observed disruption of parasympathetic influences on the heart was consistent with predictions and suggests that an index of vagal tone may be combined with existing measures to improve the accuracy of computer decisions. However, we did not anticipate that the effects on vagal tone would occur as late as 15 seconds after stimulus onset. Since we had programmed a poststimulus onset recording interval of only 20 seconds, we may have attenuated the effects on vagal tone. The present study assessed the reliability of our previous finding and tested the possibility that stronger effects would be observed if the poststimulus onset recording interval were extended to 25 seconds.

Another measure that merits additional research is arterial blood pressure. For years it has been assumed that the cardiograph provides an indirect measure of blood pressure (Geddes & Newberg, 1977). However, results from two recent studies have been mixed. In each of these studies, absolute blood pressure was recorded continuously from a Finapres arterial pressure monitor. The Finapres uses a finger cuff and servo-mechanism to maintain a constant pressure in the cuff. As blood pressure in the finger changes, the servo-system rapidly adjusts the cuff pressure to compensate. The changes in cuff pressure are then transduced and transformed into continuous measures of blood pressure. As compared to the traditional cardiograph cuff, the Finapres cuff may be inflated for long periods of time with little or no discomfort, is considerably less invasive, and is less sensitive to movement artifact.

In one study, Numaguchi, Kircher, Raskin, Packard, and Woltz (1994) recorded the cardiograph and blood pressure with the Finapres while individuals performed mental arithmetic, valsalva maneuvers, and a numbers test. They computed within-subject cross-lagged correlations between time series of cardiographic and Finapres measures of cardiovascular activity. Changes in the baseline of the cardiograph did not covary with systolic or diastolic blood pressure. However, changes in the amplitude of cardiograph pulses were significantly correlated with both systolic and diastolic blood pressure. A decrease in the amplitude of cardiograph pulses was most strongly associated with increases in diastolic blood pressure.

After the Numaguchi et al. study, Podlesny and Kircher (1999) conducted a mock crime experiment to explore the possibility of replacing the cardiograph with the Finapres in probable-lie polygraph examinations. In contrast to results obtained by Numaguchi et al., baseline changes in the cardiograph were highly correlated with diastolic blood pressure (r > .80) and were somewhat less highly correlated with systolic blood pressure. Interestingly, however, peak increases in systolic blood pressure were more diagnostic of the truthful/deceptive criterion (r_{pb} = .53) than were peak increases in diastolic blood pressure (r_{pb} = .43) or the cardiograph (r_{pb} = .39). These findings suggest that there may be some advantage in abandoning the cardiograph in favor of the Finapres or some similar device. The present study assessed the reliability of these findings and tested if the relative superiority of the systolic measure is maintained in tests with directed-lie questions.

Finally, this research project allowed us to re-examine the respiration responses of examinees during probable-lie and directed-lie examinations. In laboratory and field studies of the probable-lie test, innocent examinees showed greater suppression in respiratory activity in response to probable-lie questions than to relevant questions, or they showed little difference in their respiration responses to probable-lie and relevant questions. In contrast, guilty examinees showed greater suppression to relevant questions (e.g., Dawson, 1981; Gatchel, Smith, & Kaplan, 1984; Kircher & Raskin, 1988; Podlesny & Raskin, 1977; Podlesny & Truslow, 1993; Raskin et al., 1988). Respiration suppression is a component of the orienting reflex (Lynn, 1966) and is consistent with the hypothesis that the test question constitutes a psychological challenge to the examinee. However, in the directed-lie tests administered by Horowitz et al. (1997), only guilty examinees exhibited the predicted pattern of respiration responses to directed-lie and relevant questions. Contrary to expectations, innocent examinees showed significantly greater respiratory suppression in response to relevant questions than to directed-lie questions. If a similar pattern of results is found in the current study, the rules of numerical scoring and computer models for diagnosing truth and deception, which have been based on years of research with probable-lie tests, should be modified if they are to be used with directed-lie tests.

Methods

Participants

Four-hundred-and-seventeen adults were recruited from the general community by newspaper advertisements. The advertisements offered $30 for two hours of participation and the opportunity to earn an additional $50 bonus. Of the 417 individuals, 81 were eliminated from the study for a variety of reasons. Thirty-three individuals assigned to the guilty condition (16%) declined to participate after they received their instructions to commit the simulated theft. Eighteen individuals failed to follow instructions (e.g., did not commit the theft yet reported for their polygraph, arrived late, or brought a child with them to the lab). Thirteen individuals were dismissed due to health problems, which included reports of pain, less than 4 hours of sleep, and high blood pressure. Nine individuals assigned to the guilty condition (5%) confessed. Equipment problems and experimenter errors resulted in the loss of eight other individuals. The remaining 168 innocent and 168 guilty participants were retained to fill the cells of the design matrix (described below).

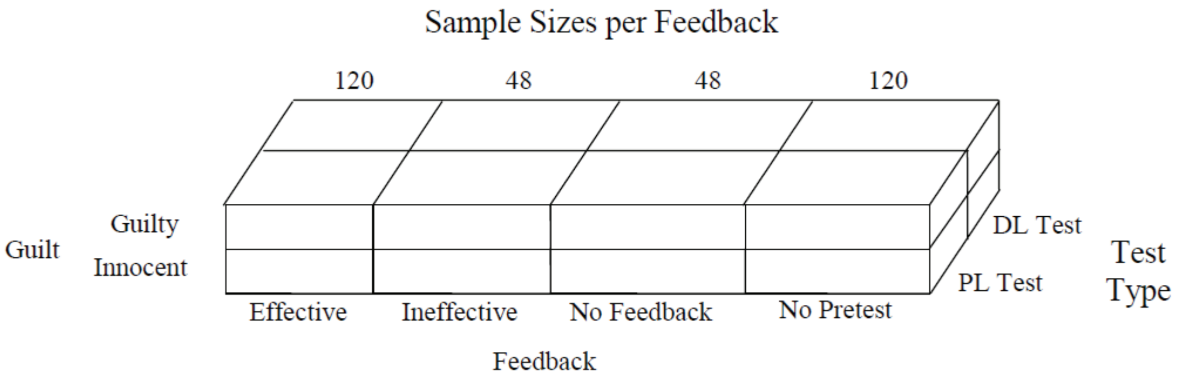
The mean age of the sample was 30.7 years (SD = 11). Years of education ranged from 9 to 25 (M = 14.3, SD = 2.5). Most participants were Caucasian (87.5%), 5.7% were Hispanic, and the remaining 6.8% were Black, Asian, or American Indian. Fifty-three percent of the participants were single, 33.9% were married, and the remaining 12.2% were divorced, separated, or widowed. Although a wide range of occupations was represented, over 75% of the sample fell into one of the following eight categories: student (17%), professional (11.9%), sales worker (9.2%), office worker (8.3%), service worker (8.3%), unemployed (7.7%), homemaker (7.7%), or laborer (7.4%).

Design

Guilty and innocent participants were randomly assigned to one of 16 cells in a completely crossed 2 x 2 x 4 factorial design with equal numbers of male and female participants in each cell. The design is illustrated in Figure 1. All factors except Sex are represented in the figure.

The first factor, Guilt, had two levels; 168 participants were guilty of committing a mock crime and the remaining 168 were innocent of the crime. The second factor, Test Type, also had two levels; half of the participants were given probable-lie comparison question tests (PL) and half were given directed-lie tests (DL).

Figure 1. Design of experiment.



The third factor, Feedback, had four levels. Participants were unevenly distributed over the four levels of the Feedback factor. One group of 120 participants (30 participants in each of the four cells shown on the far left of Figure 1) received the type of feedback commonly provided to suspects in actual field examinations. Prior to their polygraph test, they were given a demonstration test and told, regardless of the outcome, that they showed their strongest reaction to the number they had chosen. They also were told they should have no problem passing the polygraph test if they answer all of the questions truthfully (effective-feedback group).

Twelve participants were assigned to each of the four ineffective-feedback cells of the design matrix. Participants who received ineffective feedback were given a numbers test and were told, regardless of the outcome, that they did not react appropriately to the chosen number. They also were told that it would be difficult to determine if they were lying or telling the truth during their polygraph test.

Thirty participants were assigned to each of the four no-pretest control groups illustrated on the far right of Figure 1. The procedures for participants in the control groups were the same as those used for other subjects except that control subjects were not given the numbers test.

To summarize, 120 participants were given the preliminary stimulation test and received feedback that the polygraph was effective. Forty-eight participants were given the pretest and received feedback that the test was ineffective. Another 48 participants were given the pretest and received no feedback. The remaining 120 participants were not given the pretest or any form of feedback. Within each level of the Feedback factor, the design was balanced in terms of numbers of guilty and innocent male and female subjects who were given either probable-lie or directed-lie polygraph examinations.

Two examiners administered all of the polygraph tests. One examiner was an advanced doctoral student in educational psychology. The graduate student (PCB) tested 12 subjects in each of the 16 cells in the design matrix (192 subjects). The remaining 144 subjects were tested by the post-doctoral research associate (BGB).

Procedures

The procedures were similar to those described elsewhere (Kircher & Raskin, 1988). Prospective participants called a secretary who screened the participants for eligibility and briefly described the experiment and pay policy. Callers were invited to participate if they met the following criteria: (1) they were between 18 and 65, (2) they were not taking prescription medication, (3) they had never had a polygraph test, (4) they were fluent in English, and (5) they had no major medical problems.

Callers who agreed to participate were given an appointment to report to a room in a building on the campus of the University of Utah. When the participant arrived, an envelope addressed to the participant was taped to the door. Instructions within the envelope directed the participant to enter the room, close the door, read and sign an informed consent form, complete a brief questionnaire, and then play a cassette recorder that presented their instructions over headphones.

Guilty participants received tape-recorded instructions to commit a mock theft of a $20 bill from a wallet that was located in a purse in a desk in a secretary's office. Participants went to a secretary's office on a different floor of the building and asked the secretary where Dr. Mitchell's office was located. The secretary was actually a confederate in the experiment. The secretary responded that there was no Dr. Mitchell in the department. The participant thanked the secretary and left the office. The participant then waited in the hallway until the secretary left the office unattended (1-3 minutes), entered the office, searched the desk for the purse, and took the $20 bill from the wallet that was in the purse. Participants were instructed to conceal the $20 on their person and go to a room where they waited for the polygraph examiner. Guilty participants were instructed to prepare an alibi in case they were caught in the office. Innocent participants listened to a general description of the crime, left the area for 15 minutes, and went to a room where they waited for the polygraph examiner.

All participants were told that a polygraph expert who did not know if they had committed the theft would give them a polygraph test. They were told that the examiner would use a computer to assist in the analysis of their polygraph charts, and if they could convince the polygraph examiner of their innocence, they would receive $80. They were also told that if they failed to convince the examiner of their innocence, they would receive only $30.

After the participant had reported to the waiting room, the polygraph examiner went to the room, introduced himself, and instructed the participant to go to the restroom and wash their hands with soap and warm water. The participant was then escorted to the lab where the test was administered. The session was videotaped and audiotaped.

Standard field polygraph procedures were used. The polygraph examiner asked about the participant's prior experiences with the polygraph and had the participant sign a Polygraph Informed Consent form. The examiner then obtained some biographical information and asked some questions about their health. Participants who reported less than 4 hours of sleep, were experiencing pain, or indicated that they had recently taken stimulant or depressant drugs (prescription or otherwise) were not tested; they were paid for their partial participation and released. For all remaining participants, the sensors were attached and adjusted to ensure adequate recordings. The examiner then described the role of the autonomic nervous system in the detection of deception.

For participants in the effective, ineffective, and no-feedback conditions, the numbers test was then conducted. The participant was told that a preliminary test with numbers would be conducted to give the participant an opportunity to practice answering questions and to give the examiner an opportunity to adjust the instrument to ensure adequate recordings. Participants were then asked to choose a number between 3 and 6 and to tell the examiner which number they had chosen. The examiner told the participant that he would ask, "Regarding the number that you chose, was it the number 1?"... "Was it the number 2?" … and so on to the number 7. The participant was told to answer "No" to each question. Participants in the effective and ineffective feedback conditions then were told that their "No" responses to all of the questions would be truthful except when asked about the number they chose. That way the examiner would be able to see what it looked like when the participant lied and when the participant told the truth. The examiner said that he also would be able to see if the participant was a suitable subject for a polygraph test.

Effective feedback. Regardless of the actual outcome, after the demonstration test, participants in the effective feedback condition were told that they showed a strong and clear reaction to when they lied and little or no reaction when they were truthful. The participant was told that, based on the results of the numbers test, they should have no problem passing the polygraph test as long as they are completely truthful to all of the questions on the test.

Ineffective feedback. Regardless of the actual outcome, after the demonstration test, participants in the ineffective feedback condition were told that it was not possible to tell when they were truthful or deceptive during the numbers test. Based on the results of the numbers test, participants were told that they were not suitable subjects for a polygraph test and the examiner probably would be unable to tell if the participant took the $20.

No-feedback. After the demonstration test, participants in the no feedback condition were told only that the examiner was able to obtain good recordings. If the participant asked about the outcome, the examiner said that he would have to do an analysis of the recordings and could do so at the end of the test if the participant was still curious about the result.
After the demonstration test, the examiner reviewed the appropriate set of questions with the participant. The test questions for participants assigned to the probable-lie condition were as follows:

(Outside Issue) 1. Do you understand that I will ask only the questions we have discussed?

(Sacrifice Relevant) 2. Do you intend to answer truthfully all of the questions about the theft of the $20?

(Neutral) 3. Do you live in the United States?

(Probable-lie) 4. Before the age of __, did you ever take something that didn't belong to you?

(Relevant) 5. Did you take that $20 from the purse?

(Neutral) 6. Is today ___?

(Probable-lie) 7. During the first __ years of your life, did you ever do anything that was dishonest or illegal?

(Relevant) 8. Did you take that $20?

(Neutral) 9. Is your first name ___?

(Probable-lie) 10. Between the ages of __ and __, did you ever lie to get out of trouble?

(Relevant) 11. Do you have that $20 with you now?

Relevant questions that pertained to the theft and the sacrifice relevant question were reviewed first, probable-lie or directed-lie comparison questions were reviewed next, and the neutral and outside issue questions were reviewed last. When the examiner introduced the probable-lie and their purpose was to determine if they were the type of person who would take something then lie about it. If the participant answered “Yes” to a probable-lie question, the question was reworded slightly to elicit a “No” response from the participant; e.g., “Other than what you told me, before the age of __, did you ever take something that didn't belong to you?

The test questions for participants assigned to the directed-lie condition were the same as those presented to participants in the probable-lie condition, except that the probable-lie questions in positions 4, 7, and 10 were replaced with the following directed-lie questions.

(Directed-lie) 4. In your entire life, did you ever tell even one lie?

(Directed-lie) 7. Have you ever broken a rule or regulation?

(Directed-lie) 10. Did you ever make a mistake?

Participants were told to lie to these questions and to answer them “No.” They also were told that it was very important that they appear to be lying to the directed-lie questions. The examiner told participants that he would not want to make a mistake and conclude that they had lied about the theft of the $20 if they were actually telling the truth, simply because they did not respond appropriately to the directed-lie questions.

The probable-lie or directed-lie test was then administered. The interval between stimulus onsets was a minimum of 25s, and the interval between charts was between one and three minutes. After the first chart, probable-lie participants were asked if there were any problems with any of the questions. After the second chart, they were asked if they felt anything unusual when they were asked one of the probable-lie questions. Directed-lie participants were asked after each chart if they were lying to the directed-lie questions and if they felt any differently when they lied. These procedures were designed to draw the participant's attention to the comparison questions and reduce the risk of false positive errors.

The question sequence was presented five times. Neutral and comparison questions were rotated over repeated presentations of the question sequence such that each relevant question was preceded by each neutral and each comparison question at least once. The orders of presentation of the questions were not reviewed with the participant in advance.

At the conclusion of the test, the sensors were removed, and the subject was asked to complete posttest questionnaires. The probability that the participant was truthful was then computed using algorithms described elsewhere (Kircher & Raskin, 1988). If the probability of truthfulness exceeded 0.70, the participant was paid $80, $30 for their time and a $50 bonus. Otherwise, the participant was paid $30. After computing a decision, the original examiner was informed by the secretary of the participant's guilt status. Participants were then debriefed and released.

Apparatus

The CPS-LAB system (Scientific Assessment Technologies, SLC, UT) was used to configure the data collection hardware, specify storage rates for the physiological signals, and build automated data collection protocols. CPS-LAB also was used to collect, edit, and score the physiological data.

The physiological data acquisition subsystem (PDAS) of CPS-LAB generated analog signals for thoracic and abdominal respiration, skin conductance, cardiograph, finger pulse amplitude, skin potential, and cardiotachometer. In addition, calibrated analog output from an Ohmeda 2300 Blood Pressure Monitor was routed to a general-purpose coupler on the PDAS. Each of the eight analog signals was digitized at 1000 Hz with a Metrabyte DAS 16F analog-to-digital converter installed in a 50 MHz PC compatible 486 computer with 16 MB of RAM.

Respiration was recorded from two Hg strain gauges secured with Velcro straps around the upper chest and the abdomen just below the rib cage. The strain gauge changed in resistance as the subject breathed. Resistance changes were recorded DC-coupled with a 2-pole, low-pass filter, f_c = 13Hz.

Skin conductance (SC) was obtained by applying a constant voltage of .5V to two UFI 10mm Ag-AgCl electrodes filled with .05M NaCl in a Unibase medium. The electrodes were taped with adhesive collars to the distal phalanx of the ring and last fingers of the left hand. The signal was recorded DC-coupled with a 2-pole, low-pass filter, f_c = 6 Hz.

The cardiograph was recorded from a blood pressure cuff wrapped around the right upper arm and inflated to 55 to 60 mm of Hg at the beginning of each chart. The cuff was connected by rubber tubing to a Motorola MPX10DP pressure transducer in the PDAS. The output from the pressure transducer was amplified and recorded DC-coupled with a 2-pole, low-pass filter, f_c = 8.8 Hz.

Finger pulse amplitude (FPA) was obtained from a UFI photoplethysmograph attached to the first finger of the left hand with a Velcro strap. The signal from the photocell was AC-coupled with a 0.2-second time constant and a 2-pole, low-pass filter, f_c = 10 Hz.

The electrocardiogram was obtained from Lead II (right arm and left leg) using disposable, pre-gelled Red Dot^{tm} Ag-AgCl snap electrodes. The ground was placed on the left upper arm. The PDAS generated a 20 ms square wave pulse that coincided with the R-wave in the electrocardiogram. The square wave from the PDAS was routed to the analog-to-digital converter, and the CPS-LAB software measured and stored the time between successsive pulses to the nearest ms (heart period).

Skin potential (SP) was recorded from Beckman 10mm Ag-AgCl electrodes filled with .05 M NaCl in a Unibase medium attached to the thumb of the left hand (active site) and the lower arm, just below the elbow (inactive site). The inactive site was rubbed with alcohol prior to applying the electrode. Skin potential was recorded DC-coupled with a 2-pole, low-pass filter, f_c = 10 Hz. A 39.2 K ohm resistor was soldered in series with the reference (inactive) electrode to prevent variations in skin potential from affecting the skin conductance recordings.

The finger cuff of the Finapres blood pressure monitor was attached with Velcro to the middle phalanx of the middle finger on the left hand. Continuous calibrated voltage changes from the Finapres monitor were routed to a general purpose coupler on the PDAS where it was recorded DC-coupled with a 2-pole, low-pass filter, f_c = 10 Hz. The voltage changes were converted to absolute blood pressure (BP) in mm of Hg.

The data for each channel were collected at 1000 Hz. The 1000 Hz samples were reduced prior to storing them on the hard disk by averaging successive sample points. Respiration and electrodermal channels were stored at 10 Hz. Cardiograph, finger pulse, and BP signals were stored at 100 Hz. The cardiotachometer produced an interbeat interval measured to the nearest ms for each heart beat.

Numerical Evaluations

The polygraph charts were scored twice, once by the independent evaluator and again by the original examiner. The independent evaluator was unaware of the participant's guilt or innocence and had had no contact with the examinee. As described above, the original examiner had used the computer to edit artifacts from the recordings and to analyze the polygraph charts for a decision. If the computer outcome was 'truthful,' the participant was paid $80. If the computer outcome was 'inconclusive' or 'deceptive,' the participant was paid only $30. After computing a decision and communicating that decision to the participant, the original examiner was informed about the participant's actual guilt or innocence. At that point, the original examiner had not yet numerically scored the charts. To reduce the effects of this knowledge on the original examiner's subsequent numerical evaluations, the original examiner waited a minimum of two weeks to score the charts. In most cases, the original examiner evaluated the charts several months after the test. Since the examiners conducted tests daily, they were not likely to associate a particular set of charts with a particular subject and treatment condition. Indeed, both examiners reported that they could not recall if the charts they were scoring had been obtained from a guilty or innocent subject. Nevertheless, since the original examiner had had contact with the subject and had learned of the subject's deceptive status prior to scoring the charts, the numerical evaluations by the original examiner were used only to assess interrater reliability. All other analyses of numerical scores and outcomes were based on the evaluations performed by the independent rater. Interrater reliability and intermethod (human and computer) correlations are reported in Appendix A.

The Utah numerical scoring system was used to assign scores and is described elsewhere (Bell, Raskin, Honts, and Kircher, 1999). Briefly, a score that ranged from -3 to +3 was assigned to respiration, SC, cardiograph, and peripheral vasomotor activity channels for each presentation of a relevant question. Each score was based on a comparison of the participant's reaction to a comparison and relevant question. A positive numerical score was assigned when the reaction to the comparison question was greater than the reaction to the adjacent relevant question. A negative score was assigned when the reaction to the relevant question was greater. A score of zero was assigned if there was little or no difference in the size of the reactions to the two types of questions. A score of 1 was assigned when the difference was small but noticeable. A score of 2 was assigned if the difference was large, and a score of 3 was assigned if the difference was dramatic, and the stronger response was the largest on the chart for that physiological measure.

A respiration reaction was indicated by a reduction in the amplitude of respiration cycles, an increase in cycle time, and/or a rise in the respiration baseline. In the present study, the magnitude of the SC response was indicated exclusively by its amplitude. The strength of the cardiograph response was indicated by its amplitude and duration. The magnitude of the vasomotor response was based primarily on the duration of the reduction in FPA and secondarily on the change from the largest to the smallest finger pulse.

The numerical scores were then summed across comparison/relevant question pairs and across the four physiological measures. The participant was reported as deceptive if the total score was –6 or less. The participant was reported as truthful if the total was +6 or higher. Scores between +5 and -5 were considered inconclusive. In accordance with standard field practice, if the test was inconclusive after the third chart, the numerical scores assigned to the fourth and fifth charts were added to the 3-chart total. The test was inconclusive only when the 5-chart total also failed to reach a +6 or –6 cutoff.

Computer Measurements of Autonomic Activity

Response curves. From the series of digitized polygraph signals, response curves were generated for SC, cardiograph, BP, peripheral vasomotor activity (FPA), and heart period. The SC response curve was defined by the series of stored samples. The heart period response curve was also defined by the series of interbeat intervals. The generation of response curves for the cardiograph, blood pressure, and vasomotor channels was more involved. The computer identified the time and level of each systolic and each diastolic point for the 20 seconds that followed stimulus onset. The computer then calculated a weighted average of the systolic points that occurred during each poststimulus second (Kircher & Raskin, 1988). The resulting time series of 20 systolic levels defined a systolic response curve. The same procedure was used to create a second-by-second diastolic response curve. For blood pressure, the systolic and diastolic response curves were analyzed separately. For the cardiograph, a mean second-by-second response curve was computed by averaging the systolic and diastolic levels for each second.

To track changes in peripheral vasomotor activity, systolic and diastolic points were identified during the interval that began two seconds prior to stimulus onset and ended 20 seconds after stimulus onset. An FPA response curve was computed by subtracting the diastolic level for each second from the corresponding systolic level. Each difference was the amplitude of the finger pulses for a given second. Each poststimulus amplitude was then divided by the mean prestimulus amplitude to obtain a 20-second time series of proportions. Physiological arousal in this measure was indicated by reductions in pulse amplitude associated with peripheral vasoconstriction. Since the scoring algorithm treated increases rather than decreases in the response curve as indications of increased arousal, the vasomotor response curve was reflected (flipped upside down) by multiplying each poststimulus proportion by –1 prior to feature extraction. In that way, a decrease in the amplitude of finger pulses was associated with a rise in the response curve.

Feature Extraction. The features extracted from response curves were as follows:

> *Peak amplitude*. Low points in the response curve were identified as changes from negative or zero slope to positive slope, and high points in the response curve were identified as changes from positive slope to zero or negative slope. The difference between each low point and every succeeding high point was computed. Peak amplitude was defined as the greatest such difference if it exceeded some preset minimum. For skin conductance, this minimum was 0.02 µ Siemens. For all other waveforms, the minimum was zero.
>
> *Area* was area under the response curve from response onset to the time at which the tracing recovered to the level at response onset or to the end of the scoring window, whichever occurred

first (time of full recovery). Response onset was defined as the low point from which peak amplitude was measured.

Latency was the time in ms from stimulus onset to response onset.

Rise time was the time in ms from response onset to the time of peak amplitude.

Recovery time was the difference in ms between the time of peak amplitude and the time at which the tracing recovered to the level at response onset or to the end of the scoring window, whichever occurred first.

Rise rate was the linear rate of increase from response onset to peak amplitude.

Recovery rate was the linear rate of decrease from peak amplitude to the time at which the tracing recovered to the level at response onset or to the end of the scoring window, whichever occurred first.

Excursion was the sum of absolute deviations between adjacent samples in the scoring window.

Variance was the variance of samples that defined the response curve.

Indices of differential reactivity to comparison and relevant questions. Comparison question techniques predict that innocent subjects will respond more strongly to comparison questions than to relevant questions, whereas guilty subjects will respond more strongly to relevant questions. Following our standard procedure (Kircher & Raskin, 1988), an index of differential reactivity to comparison and relevant questions was computed for each subject and each autonomic measure. For example, each subject provided 18 measurements of SC amplitude for the three comparison questions and the three relevant questions on each of the first three charts. The 18 measurements were converted to z scores. The mean of the nine z scores for relevant questions was then subtracted from the mean of the nine z scores for comparison questions.

An index of differential reactivity is analogous to the total numerical score assigned by the polygraph examiner for a particular channel. The index was positive when the mean reaction to comparison questions was greater than the mean reaction to relevant questions, and the index was negative when the reactions to relevant questions were greater. Since innocent subjects were expected to react more strongly to comparison questions and guilty subjects were expected to react more strongly to relevant questions, we expected positive scores for innocent subjects and negative scores for guilty subjects.

For all variables except respiration, a large measured response was indicative of a strong reaction. For respiration excursion, suppressed respiratory activity was indicative of a strong reaction. Thus, innocent subjects were expected to show relatively small measured respiration responses (suppression) to comparison questions, whereas guilty subjects were expected to show relatively small measured respiration responses (suppression) to relevant questions. To achieve a common direction for predicted effects, the sign of the index of differential reactivity for respiration was reversed.

Computer Decisions

Indices of differential reactivity for respiration excursion, SC amplitude, and cardiograph amplitude (baseline increase) were weighed and summed to obtain a discriminant score for each participant (Kircher & Raskin, 1988; Podlesny & Kircher, 1999). The weights for variables in the discriminant function were empirically-derived from prior samples of laboratory and field cases (Raskin et al., 1988). The discriminant score was analogous to the total numerical score assigned by a polygraph interpreter. The discriminant score for an individual was used in combination with distributions of discriminant scores for known truthful and deceptive subjects to compute the probability of truthfulness.

If the probability of truthfulness based on the first three charts of physiological data was .70 or greater, the individual was classified as truthful. If the probability of truthfulness after three charts was .30 or less, the individual was classified as deceptive. If the probability fell between those two cutoffs, a new discriminant score was computed based on all five polygraph charts. If the probability of truthfulness based on five charts exceeded a .70 or .30 cutoff, the individual was classified as truthful or deceptive. Otherwise, the test was considered inconclusive.

Results and Discussion

The Results and Discussion section begins with tables that summarize the outcomes of numerical and computer evaluations of the physiological recordings for each of the 16 cells of the design matrix. The remainder of the report is then organized into four distinct sections or "studies." The results of each study are presented and discussed before those of the next study. Study 1 focuses on the effects of the demonstration pretest and feedback to the subject about the outcome of the pretest. Study 2 compares outcomes from PL and DL tests. Study 3 reports on the reliability and validity of new physiological measures and computer decision models for PL and DL tests. Study 4 focuses on the relationship between personality measures and polygraph outcomes. Preliminary analyses that included tests for effects of treatment-related attrition and Sex are reported Appendix B. The report concludes with a general summary and set of recommendations.

Numerical Decisions

The percentages of correct, wrong, and inconclusive outcomes for independent numerical evaluations are presented in Table 1 for each treatment condition. Across all conditions, 67% of independent numerical decisions were correct, 12% were wrong, and 21% were inconclusive. Excluding inconclusive outcomes, 85% of definite decisions were correct.

Table 1. Percent Correct (C), Wrong (W), Inconclusive (I), and Correct Decisions excluding inconclusives (CD) for independent numerical decisions.

		Probable-Lie					Directed-Lie				
		B	C	W	I	CD	B	C	W	I	CD
No Pretest	Innocent[b]	30	37	33	30	52	30	77	3	20	96
	Guilty[a]	30	80	3	17	96	30	57	17	27	77
Effective Feedback	Innocent[a,b]	30	70	13	17	84	30	73	13	13	85
	Guilty[a]	30	87	3	10	96	30	67	13	20	83
No Feedback	Innocent[b]	12	75	0	25	100	12	83	0	17	100
	Guilty[a,c]	12	92	0	8	100	12	8	42	50	17
Ineffective Feedback	Innocent[a]	12	83	8	8	91	12	67	0	33	100
	Guilty	12	58	0	42	100	12	33	33	33	50

[a] The percentage of correct classifications (C) exceeded 50% for the subjects who received probable-lie tests.
[b] The percentage of correct classifications (C) exceeded 50% for the subjects who received directed-lie tests.
[c] The percentage of correct classifications (C) was significantly *less than* 50% for the subjects who received directed-lie tests..

Binomial tests were conducted to determine if the percentage of participants correctly classified as innocent or guilty exceeded chance (50%). The effective-feedback and no-pretest control conditions each contained 30 participants. For cells with 30 participants, the percentage correct (C) had to exceed 67% to achieve statistical significance (α = .05, two-tailed). The percent correct exceeded 67% in all of the effective-feedback groups except for guilty participants who received DL tests. The percent correct was also significant for guilty PL participants who received no pretest and for innocent DL participants who received no pretest.

Binomial tests of the percentage correct against 50% treated all outcomes that were not correct as errors. Since inconclusive outcomes were not correct, they were treated as errors. Despite the loss of power, when inconclusive outcomes were excluded, decision accuracy for all effective-feedback and no-pretest groups exceeded 50% (chance), except for innocent PL participants who received no pretest.

Table 1 also presents the results of binomial tests for cells with 12 participants when inconclusive outcomes were counted as incorrect. In cells with 12 cases, the power to detect effects was low. To be statistically significant, 10 of 12 subjects had to be correctly classified. In three of eight conditions, the percentage of cases classified correctly exceeded 50%, and in one condition, the percentage of cases incorrectly classified exceeded 50%. Excluding inconclusive outcomes, decision accuracy exceeded 50% for all groups except innocent PL participants in the no-pretest condition, guilty DL participants in the no-feedback condition, and guilty DL participants in ineffective-feedback condition.

Computer Decisions

The percentage of correct, wrong, and inconclusive outcomes for the computer are presented in Table 2 for each treatment condition. Across all conditions, 76% of computer decisions were correct, 16% were wrong, and 8% were inconclusive. Excluding inconclusive outcomes, 83% of definite decisions were correct.

Table 2. Percent Correct (C), Wrong (W), Inconclusive (I), and Correct Decisions excluding inconclusives (CD) for computer decisions.

		Probable-Lie					Directed-Lie				
		B	**C**	**W**	**I**	**CD**	**B**	**C**	**W**	**I**	**CD**
No Pretest	Innocent[b]	30	60	30	10	67	30	80	17	3	83
	Guilty[a]	30	73	10	17	88	30	57	27	17	68
Effective Feedback	Innocent[a,b]	30	90	7	3	93	30	83	13	3	86
	Guilty[a,b]	30	87	13	0	87	30	67	17	17	80
No Feedback	Innocent[a,b]	12	100	0	0	100	12	92	0	8	100
	Guilty[a]	12	92	0	8	100	12	33	67	0	33
Ineffective Feedback	Innocent[a,b]	12	92	8	0	92	12	92	8	0	92
	Guilty	12	75	8	17	90	12	58	17	25	78

[a] The percentage of correct classifications (C) exceeded 50% for the subjects who received probable-lie tests.
[b] The percentage of correct classifications (C) exceeded 50% for the subjects who received directed-lie tests.

The results of binomial tests of the percent correct (C) against chance (50%) are summarized in Table 2. The percent correct exceeded chance in 11 of 16 cells. The same 11 cells yielded decision accuracies that exceeded 50% when inconclusive outcomes were excluded. Excluding inconclusive outcomes, the percentage of correct decisions was also greater than 50% for guilty PL participants in the ineffective-feedback condition (90%).

Study 1: Effects of the Demonstration Test and Feedback on Polygraph Outcomes

Analytic Plan

The analytic plan called for separate analyses of numerical and computer outcomes (correct, wrong, and inconclusive), total numerical scores, and selected computer measurements. We used chi-square tests to compare distributions of correct, wrong, and inconclusive numerical or computer outcomes. For each of the major hypotheses, four separate chi-square tests were conducted -- one for guilty PL participants, one for innocent PL participants, one for guilty DL participants, and one for innocent DL participants. Each chi-square tested a 2 X 3 matrix of outcomes. One dimension of the matrix was Outcomes with three levels (correct, wrong, and inconclusive). The other dimension was the factor of interest, and it had two levels (e.g., whether or not participants received a demonstration test).

We used the method of planned comparisons to analyze numerical scores and selected computer measurements. Each planned comparison was a simple 2 X 2 interaction contrast with 1 df. In general, we tested if the discrimination between guilty and innocent participants in one condition was greater (or less) than the discrimination between guilty and innocent participants in another condition. The advantage in using the method of planned comparisons instead of ANOVA was that the error term for the statistical test was based on all of the participants, not only those in the four treatment conditions involved in the comparison. Consequently, the method of planned comparisons had more error degrees of freedom and more power to detect simple interaction effects than would an ANOVA that included only the individuals in the four treatment groups under consideration (Keppel, 1991).

Combined Effects of the Demonstration Test and Effective

Feedback on polygraph outcomes. The primary objective of the present study was to test if the preliminary numbers test and effective feedback are necessary to achieve high levels of discrimination between truthful and deceptive individuals. To answer this question, we compared the effective-feedback condition to the no-pretest control condition. Consistent with current field practice, participants in the effective-feedback condition were given the numbers test and feedback that the polygraph was effective in distinguishing between their truthful and deceptive answers on the pretest. The dependent variables for these comparisons consisted of numerical outcomes (correct, wrong, and inconclusive), total numerical scores, computer outcomes, and three computer measurements.

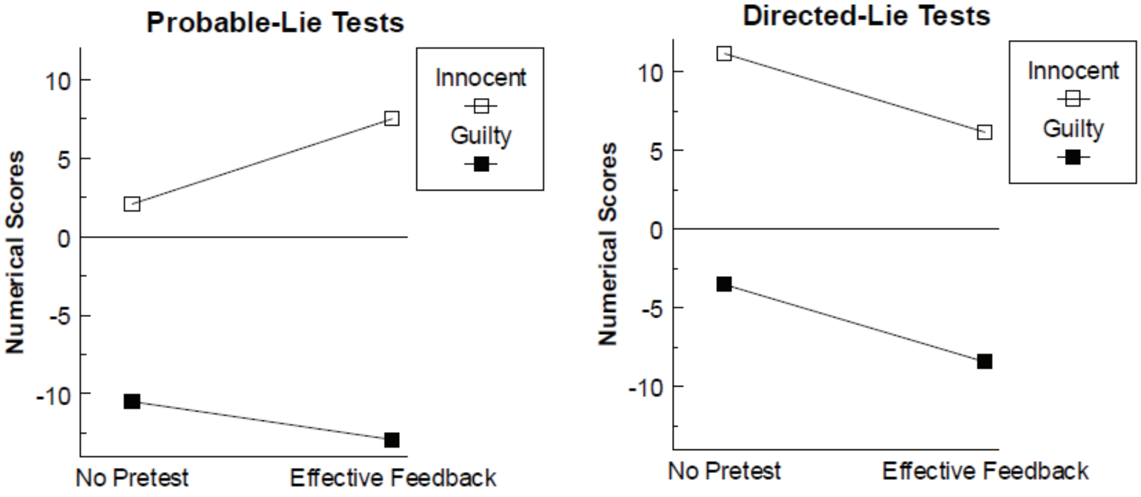
Numerical outcomes. Outcomes (correct, wrong, and inconclusive) for participants who received no pretest were compared to outcomes for participants who received the numbers pretest and effective feedback. The data for these analyses are present in rows 1 through 4 in Table 1 above.

For the PL test, effective-feedback resulted in fewer false positive errors and fewer inconclusive outcomes for innocent participants, χ^2 (2) = 6.84, p < .05. Although more guilty PL participants were correctly classified in the effective-feedback (87%) than in the no-pretest condition (80%), the benefits of the pretest for guilty participants were not significant. The pretest and effective feedback had no significant effect on the outcomes for innocent or guilty participants who were given DL tests.

Numerical scores. A planned 2 X 2 interaction contrast was performed by comparing the guilty and innocent group means for the effective-feedback condition to the guilty and innocent group means for the no-pretest condition. The predicted interaction effect on total numerical scores was significant for the PL test, t(320) = 2.35, p < .05, but not for the DL test. The means for the PL and DL tests are plotted in Figure 2. Examination of Figure 2 reveals that the PL test yielded better discrimination between guilty and innocent participants when it was preceded by the demonstration test and effective feedback than when it was not. For the DL test, the pretest and effective feedback shifted the numerical scores in the negative direction and balanced the percentage of false positive and false negative numerical decision errors.

Computer outcomes. To assess the combined effects of the pretest and effective feedback on computer decisions, the outcomes (correct, wrong, and inconclusive) for the effective-feedback condition were compared to the outcomes for the no-pretest condition. Separate tests were performed for innocent and guilty participants who received either PL or DL tests.

Figure 2. Mean total independent numerical scores for the no-pretest (n = 30/group) and effective-feedback (n = 30/group) conditions

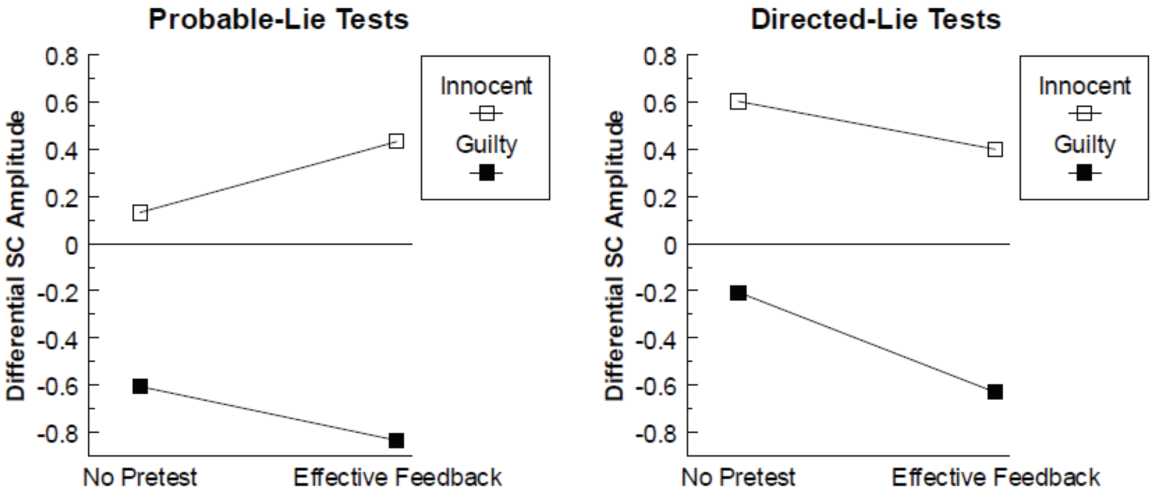


Examination of rows 1 through 4 of Table 2 reveals that decision accuracy tended to be higher for participants in the effective-feedback conditions than in the corresponding no-pretest conditions. Statistical analysis confirmed that there were significantly fewer false positive errors and fewer inconclusive outcomes for innocent PL participants in the effective-feedback condition than for innocent PL participants in the no-pretest condition, χ^2 (2) = 7.26, p < .05. However, the benefits of the demonstration test and effective feedback were not significant for guilty PL participants, innocent DL participants, or guilty DL participants. The same pattern of results was observed for numerical outcomes.

Computer measurements. Planned 2 X 2 interaction contrasts were conducted to assess the combined effects of the demonstration pretest and effective feedback on SC amplitude, cardiograph amplitude, and respiration excursion. The effect on SC amplitude was significant for the PL test, t(320) = 2.46, p < .05, but not for the DL test. For the PL test, discrimination between guilty and innocent groups on measures of SC amplitude was greater in the effective-feedback condition than in the no-pretest condition. The interaction effect on SC amplitude is illustrated in Figure 3. The pattern of cell means for SC amplitude was similar to that shown in Figure 2 for total numerical scores. The effects on cardiograph and respiration measures were not significant for PL or DL tests.

Participants in the no-feedback condition received a demonstration test but no feedback about the outcome. Simply having experienced a pretest in the absence of any feedback may affect polygraph outcomes because it provides an opportunity for participants to become accustomed to being interrogated while attached to a polygraph prior to taking their polygraph test. To

Figure 3. Mean indices of differential SC responses for no-pretest (n = 30/group) and effective-feedback conditions (n = 30/group)



determine if the pretest by itself affected outcomes, no-feedback participants who received the pretest but no feedback (only pretest) were compared to those who did not receive the pretest (no-pretest).

Numerical outcomes. For the PL test, there was no effect of the demonstration test on decision outcomes for guilty participants. For innocent participants, decision accuracy was significantly greater in the no-feedback (only-pretest) condition than in the no-pretest condition, χ^2 (2) = 6.72, p < .05. For the DL test, decision accuracy on guilty participants was significantly lower for the no-feedback group than for the no-pretest group, χ^2 (2) = 8.32, p < .05. Excluding inconclusive outcomes, decision accuracy for guilty DL participants dropped from 77% in the no-pretest condition to only 17% in the no-feedback condition. In contrast, decision accuracy for innocent DL participants was slightly, but not significantly, higher in the no-feedback condition (100%) than in the no-pretest condition (96%).

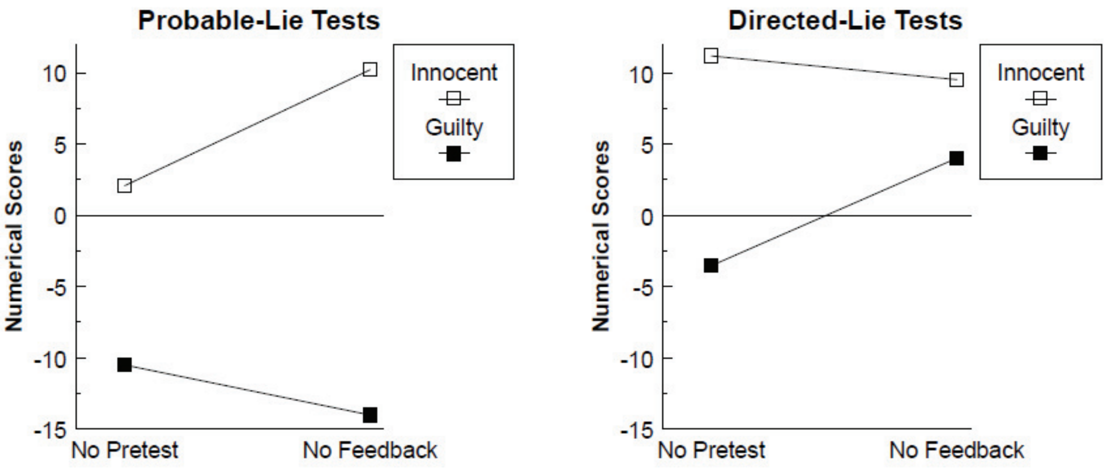
Numerical scores. For the PL test, a test of the Guilt X Demonstration Test (Yes/No) interaction was significant for total numerical scores, t(320) = 2.53, p < .01. Mean total numerical scores for the PL no-pretest and no-feedback conditions are shown in the left panel of Figure 4. Discrimination between guilty and innocent participants was greater when participants received the pretest (no-feedback) than when they did not (no-pretest).

For the DL test, the Guilt X Demonstration Test interaction was also significant, t(320) = -2.09, p< .05. However, in this case, there was less discrimination between guilty and innocent participants who received the pretest (no-feedback) than for those who did not receive the pretest. The right panel of Figure 4 shows the mean numerical scores for DL no-pretest and no-feedback conditions.

Computer decisions. For the PL test, decision accuracies for guilty and innocent participants were not significantly greater in the no-feedback condition than in the no-pretest condition. For the DL test, there was no effect of the demonstration test on innocent participants, but there was a cost associated with the demonstration test for guilty DL participants, χ^2 (2) = 8.32, p < .05.

Computer measurements. Effects of the Guilt X Demonstration Test interaction on SC amplitude, cardiograph amplitude, and respiration excursion were evaluated separately for PL and DL tests. For the PL test, the interaction effect was significant for only SC amplitude. The pattern of cell means for guilty and innocent no-pretest and no-feedback groups was very similar to that shown above in Figure 4 for PL tests. None of

Figure 4. Mean total independent numerical scores for the no-pretest (n = 30/group) and no-feedback (n = 12/group) conditions



the interaction contrasts was significant for the DL test.

Effects of Feedback about the Demonstration Test

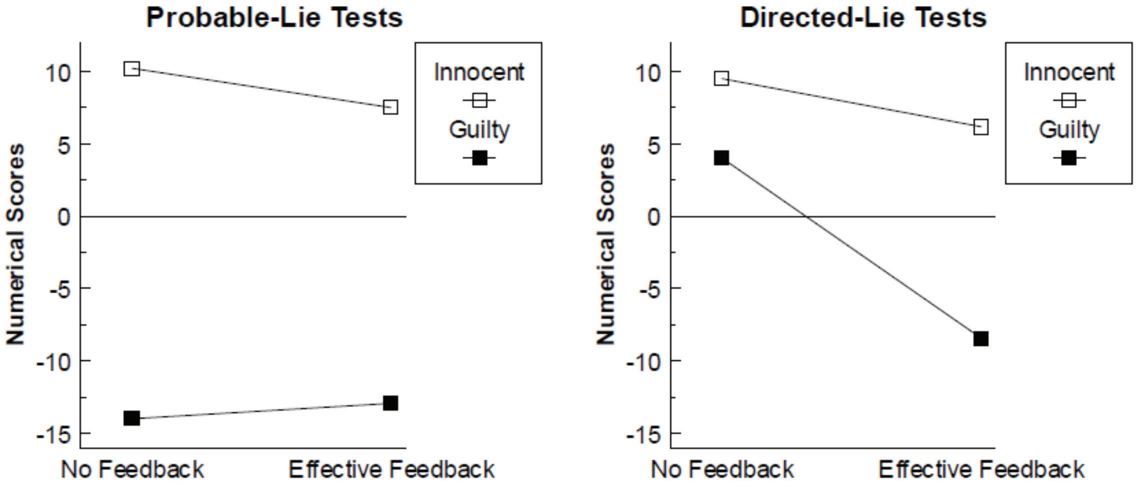
Participants in the effective-feedback, no-feedback, and ineffective feedback groups were given demonstration tests prior to their polygraph examinations. Effective-feedback participants were given feedback that the polygraph was effective; they were told that the polygraph clearly showed when they were truthful and deceptive during the demonstration test. No-feedback participants were told nothing about their responses during the demonstration test. Ineffective-feedback participants were told that the polygraph failed to indicate when they told the truth and when they lied on the demonstration test, and it would be difficult to determine if they were truthful or deceptive on the main test.

To test if feedback that the polygraph is effective affected the accuracy of the subsequent polygraph test, the effective-feedback condition was compared to the no-feedback condition. To test if feedback that the polygraph is ineffective affected the accuracy of the subsequent polygraph examination, the ineffective-feedback condition was compared to the no-feedback condition. Statistical analyses of numerical and computer outcomes were performed separately for guilty and innocent participants and for PL and DL tests.

Numerical Outcomes. Numerical outcomes (correct, wrong, and inconclusive) for the effective-feedback condition did not differ significantly from the outcomes for no-feedback condition for either the guilty or the innocent participants given PL tests. The outcomes also did not differ for innocent participants who were given DL tests. However, for guilty participants given DL tests, effective feedback was associated with fewer false negative errors and fewer inconclusive outcomes, χ^2 (2) = 11.14, p < .01. Telling participants that the demonstration test revealed their deception improved the accuracy of numerical decisions, but only for guilty participants who were given DL tests. The same pattern of results was observed when inconclusive outcomes were excluded and only definite decisions were considered. All differences in numerical decisions between ineffective-feedback and no-feedback conditions were non-significant.

Numerical Scores. Planned 2 X 2 interaction contrasts were conducted to test for effects of effective and ineffective feedback on total numerical scores. Effective feedback did not affect discrimination between guilty and innocent PL participants. However, as compared to the no-feedback condition, effective feedback improved discrimination between guilty and innocent DL participants, t(320) = 2.07, p < .05. Figure 5 shows the means for guilty and innocent participants who received either PL or DL tests. As compared to the no-feedback condition, total numerical scores for DL participants were more evenly balanced about zero in the effective-feedback condition. There were no significant effects of ineffective feedback on total numerical scores.

Figure 5. Mean total independent numerical scores for the no-feedback (n=12/group) and effective-feedback conditions (n=30/group)



Computer Outcomes. The results obtained from comparisons of computer outcomes were similar to those obtained for numerical outcomes. There were no effects of feedback on computer outcomes for guilty and innocent PL participants. For guilty DL participants, computer outcomes were significantly more accurate for those who received effective feedback than for those who did not, χ^2 (2) = 10.59, p < .01. However, computer outcomes were significantly more accurate for guilty DL participants who received ineffective-feedback than for those who did not, χ^2 (2) = 7.42, p < .05. Although the former result is consistent with predictions, the latter result certainly was not. The latter result suggests that the decision accuracy for the guilty no-feedback DL condition is spuriously low. If the present sample of guilty no-feedback DL participants is not representative of that population, then all comparisons with the guilty no-feedback DL condition are suspect and should be replicated before any serious attempt is made to interpret them.

Computer Measurements. Planned 2 X 2 interaction contrasts were conducted to assess the effects of effective and ineffective feedback on SC amplitude, cardiograph amplitude, and respiration excursion. Twelve comparisons were performed; effective-feedback was compared to no-feedback and ineffective-feedback was compared to no-feedback for each type of test (PL and DL) and for each of three physiological measures. Only one of the 12 comparisons approached statistical significance, and it included the guilty no-feedback DL participants.

Discussion

Study 1 tested if the combination of a preliminary demonstration test and effective feedback improves the accuracy of polygraph examinations. For the PL test, the percentage of correct computer decisions was 22% higher for guilty and innocent participants who received the pretest and effective feedback than for participants who received no pretest. For the DL test, the percentage of correct decisions averaged 7% higher for participants who received the pretest. Excluding inconclusive outcomes, the pretest with effective feedback increased the accuracy of decisions from 77% to 90% for the PL test and from 75% to 83% for the DL test. For the PL test, the effects of the pretest and feedback on numerical evaluations were similar to those obtained for the computer analysis. For the DL test, overall decision accuracy was about the same for the effective-feedback and no-pretest groups. However, false positive and false negative error rates for the numerical evaluator were more evenly balanced in the group that received the pretest and effective feedback. These findings are consistent with those reported by Bradley and Janisse (1981) and suggest that field polygraph examiners should continue to administer the pretest and effective feedback.

Effects of the demonstration test (pretest). Participants in the effective feedback condition differed from those in the no-pretest control condition on two dimensions; they received a pretest, and they received feedback that the polygraph accurately revealed their deception on the pretest. Since there were two factors that distinguished the effective feedback condition from the no-pretest control condition, the improvements in decision accuracy could be due to the pretest, the feedback, or both. In the absence of any feedback, the pretest allows subjects to habituate to the situation and testing procedure prior to the actual polygraph examination. Habituation and familiarization with the protocol may serve to reduce random variation in the physiological activity, enhance the signal-to-noise ratio, and improve detection rates. To explore the possibility that the pretest by itself affects decision accuracy, we compared groups that received no pretest to those that received the pretest but no feedback.

For the PL test, the pretest had a positive effect on the accuracy of decisions for innocent participants. The pretest reduced inconclusive outcomes and increased the accuracy of definite decisions from 52% to 100% for the numerical evaluator and from 67% to 100% for the computer. The pretest was also associated with a reduction in inconclusive outcomes and an increase in the accuracy of decisions on guilty participants. Excluding inconclusive outcomes, decision accuracy increased from 96% to 100% for the numerical evaluator and from 88% to 100% for the computer. Although the benefits of the pretest on decisions were not significant for guilty participants, the accuracy rate was high for guilty participants in the no-pretest condition leaving little room for improvement.

For the DL test, numerical and computer decision accuracy was about 10% higher on average for innocent DL participants in the pretest condition than in the no-pretest condition. However, the pretest was also associated with a dramatic drop in the accuracy of decisions on guilty participants. Excluding inconclusive outcomes, the accuracy of decisions for the numerical evaluator, dropped from 77% to 17%; and for the computer, accuracy dropped from 68% to 33%. Examination of the physiological measures revealed that the effects on decisions were driven by SC responses.

Why administration of the pretest with no feedback should have such an adverse effect on the accuracy of outcomes, and why the pretest would affect only guilty DL participants is a mystery. We suspect that the effect is spurious for several reasons. First, we are unaware of any theory that would predict or could account for such an effect. Secondly, because there were only 12 subjects in the guilty no-feedback condition, the effects were large but reached only the 0.05 level of significance. The probability that the result is a Type I error is small but nontrivial. We believe that larger samples of guilty and innocent DL no-feedback participants should be used to determine if the effect is dependable. Thirdly, decision accuracy was significantly lower for the guilty no-feedback group than for the guilty effective feedback and for guilty ineffective feedback groups. It is conceivable that feedback indicating the subject’s deception was clearly revealed during the pretest would improve decision accuracy. However, it is not likely that feedback that the polygraph is ineffective would also improve decision accuracy. Theoretically, the lowest level of accuracy should have been obtained from the ineffective feedback group. Since the decision accuracy was significantly lower in the no-feedback group than the ineffective feedback group, it appears that the accuracy of decisions for the guilty no-feedback group is spuriously low. Our sample of guilty DL participants who received no feedback does not appear to be representative of that particular population of guilty subjects. Additional research would be needed to test this hypothesis.

Effects of feedback. In the field, suspects who receive effective feedback are told that the polygraph accurately revealed their attempts to deceive during the pretest. Although in many cases, the examiner's feedback is correct, in other cases, the feedback is false (Horowitz, Raskin, Kircher & Honts, 1986). Since the benefits associated with the pretest and effective feedback might be due only to the pretest and not the feedback, it might not be necessary to provide subjects with any feedback about their responses during the pretest (Podlesny & Truslow, 1993).

To determine if effective feedback increases decision accuracy over and above that already afforded by the pretest, participants who received the pretest plus effective feedback were compared to participants who received only the pretest. The results of these comparisons were mixed. For the PL test, computer and numerical decisions were slightly, but not significantly, more accurate in the no-feedback condition than in the effective-feedback condition. Thus, for PL tests, it appears that effective feedback does not contribute to accurate decisions. The observed gains in accuracy, relative to the no-pretest control condition, may be attributed to the pretest and the opportunities it provides for subjects to become familiar with the monitoring equipment and their task.

By not providing feedback, polygraph examiners could avoid having to mislead some examinees. Examiners' interactions with all subjects would be more straightforward and honest. Moreover, some subjects, rightly or wrongly, may be convinced that they reacted more strongly to a question on the pretest other than the one identified by the examiner. By telling subjects that they reacted more strongly when they lied than when they told the truth, the examiner risks losing credibility with the subject. If it is not necessary to provide feedback, then several uncontrolled factors that could influence accuracy rates could be eliminated. For example, there may be differences among examiners in how convincing they are when they present the feedback. There may be differences among subjects in how they respond to that feedback. Finally, a given subject's response to the feedback may depend on who delivers it. In other words, a subject-by-examiner interaction might also affect accuracy rates. These extraneous factors and potential sources of variance would be eliminated if the polygraph examiner did not try to convince examinees that their deception was revealed by the polygraph during the pretest.

Future research might consider assessing an alternative strategy that was not evaluated in the present study. The polygraph examiner could introduce the pretest in the manner described by Podlesny and Truslow (1993). The examinee is told, quite truthfully, that the purpose of the pretest is to provide opportunities for the examiner to adjust the instrument and for the examinee to practice answering questions. If the examinee's strongest response is to the chosen number, the examiner could provide that feedback and even show the chart to the examinee. If the examinee does not show their strongest reaction to the chosen number, the examiner would simply proceed with the review of test questions and not mention the result.

The effective-feedback conditions were also compared to the no-feedback conditions for participants who were given DL tests. Effective feedback had no effect on outcomes for innocent participants, and it increased the accuracy of decisions for guilty participants. However, in light of the unusually low detection rate for guilty DL participants in the no-feedback condition, that finding was not surprising. Unfortunately, the problems discussed earlier with the guilty no-feedback DL group make any comparison with that group suspect. At this point, we do not know if the accuracy of decisions for guilty DL subjects is affected by the pretest or effective feedback. As noted above, the combined effects of the pretest and effective feedback were associated with some improvement in the numerical and computer outcomes on guilty DL participants, but the effects were not statistically significant.

As compared to effective feedback, ineffective feedback tended to increase the percentage of inconclusive outcomes for guilty participants, but it had little effect on the accuracy of decisions. Analyses of self-report data also failed to reveal effects of ineffective feedback on perceptions of polygraph accuracy. Together, the results may indicate that our manipulation was weak. Other aspects of the situation were inconsistent with the ineffective-feedback manipulation and conveyed the message that polygraph techniques are highly effective. For example, participants were told that they would be given a polygraph test by an expert polygraph examiner, and the examiner maintained a competent and professional demeanor across all treatment conditions. The tests were administered in a scientific laboratory on a large university campus. The laboratory was populated with computers, audiovisual electronics, and physiological recording equipment, all of which were in plain view of the participants as they walked to a room within the laboratory for their polygraph tests. Twelve transducers and electrodes were attached to the participants to monitor autonomic and somatic activity. In addition, participants had been informed that a computer would be used to analyze the physiological data and help make the decision. Such contextual factors may have contributed to the impression the technology for detecting deception is well developed and highly effective. This message may have competed with the negative feedback and attenuated any adverse effects of the feedback on polygraph accuracy.

In summary, the results of Study 1 suggest that the pretest should be administered prior to a PL examination, and effective feedback is unnecessary. Indeed, decisions were 100% correct for the guilty and innocent PL participants who received only the pretest and no feedback. The present findings also suggest that the pretest should also be performed prior to a DL test but only if it is accompanied by effective feedback.

Study 2: Comparisons of Probable-Lie and Directed-Lie Tests

Comparisons of PL and DL tests were limited to participants in the effective-feedback condition because that condition is most similar to current field practice.

Numerical outcomes. The outcomes of numerical evaluations from PL and DL tests did not differ for innocent participants. Although decision accuracy was somewhat higher for guilty participants in the PL condition than in the DL condition (row 4 of Table 1), the difference was not significant. Numerical outcomes for PL and DL tests did not differ whether or not inconclusive outcomes were included in the analysis.

Numerical scores. The test of the 2 X 2 interaction contrast of numerical scores from PL and DL tests was not significant, t(320) = 1.75, *p* < .09. Total numerical scores for guilty and innocent DL participants were comparable to those obtained from PL participants.

Computer outcomes. For innocent participants, computer outcomes for PL and DL tests did not differ. For guilty participants, the difference between PL and DL computer outcomes approached significance, χ^2 (2) = 5.89, *p* < .06. Excluding inconclusive outcomes, there were no reliable differences between computer outcomes for PL and DL tests for innocent or guilty participants.

Computer measurements. Planned Guilt (2) X Test Type (2) interaction contrasts were tested for SC amplitude, cardiograph amplitude, and respiration excursion. The interaction was significant for only respiration excursion, t(320) = 2.89, p < .01. The mean index of differential respiration reactivity is displayed in Figure 6 for the guilty and innocent PL and DL conditions. As expected, guilty PL participants reacted more strongly to relevant than to probable-lie questions, and innocent PL participants reacted more strongly to probable-lie questions. Guilty DL participants also responded in the expected manner to relevant and probable-lie questions, but innocent DL participants did not. The pattern of respiration responses to directed-lie and relevant questions by guilty and innocent DL participants were virtually indistinguishable. Both guilty and innocent DL participants showed a greater reduction in respiratory activity in response to relevant questions than to directed-lie questions.

Respiration responses during directed-Lie tests. To explore the patterns of respiration responses from DL tests in more detail, thoracic and abdominal respiration excursion was measured on a second-by-second basis. To correct for differences in signal gain across charts and participants, the data for each channel and chart were converted to z-scores prior to extracting measurements of excursion. Mean respiration excursion curves were then computed by averaging the second-by-second measurements of thoracic and abdominal respiration excursion. Figure 7 shows the response curves for guilty and innocent DL participants and for neutral, directed-lie, and relevant questions. Conceptually, the curves show the mean level of respiration activity over time, where relatively low scores indicate suppression. Neutral questions were included to obtain a baseline measure of respiratory activity.

Figure 6. Mean indices of differential respiration responses for probable-lie (n = 30/group) and directed-lie tests (n = 30/group)

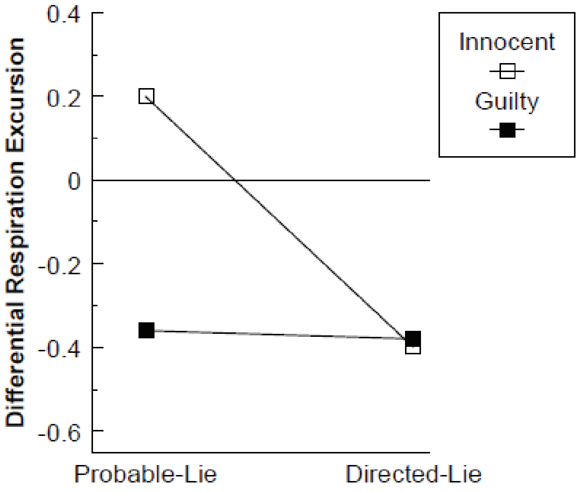


Figure 7a. Respiration excursion associated with neutral, directed-lie, and relevant questions for guilty directed-lie participants (n=30)

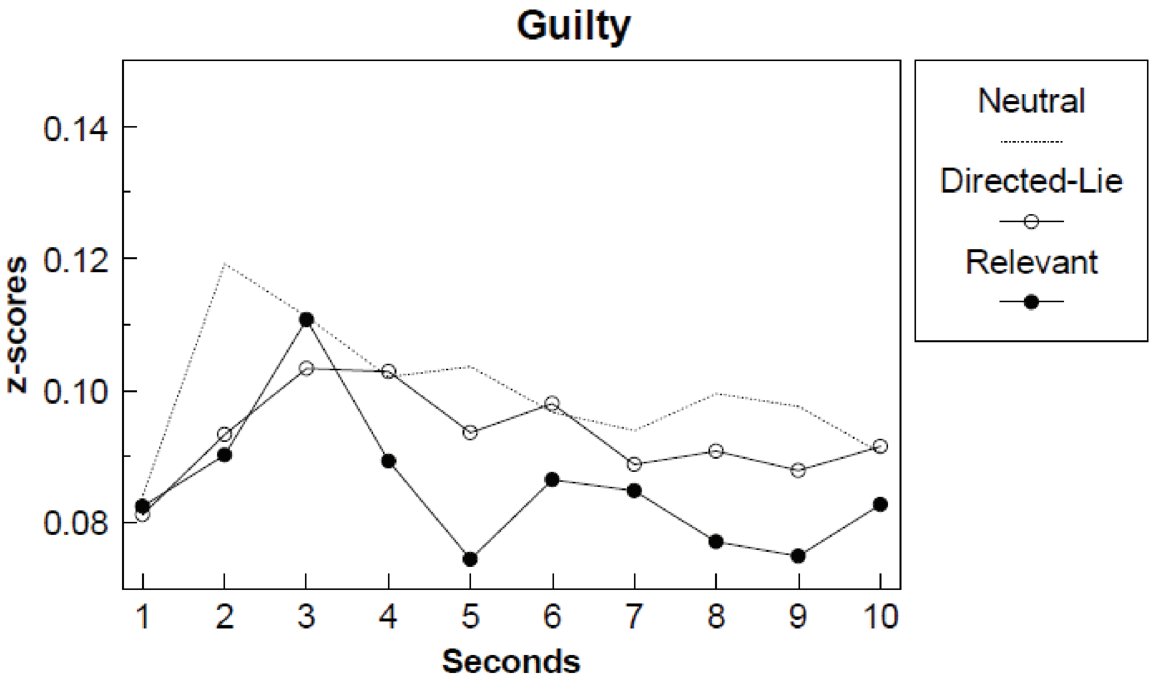
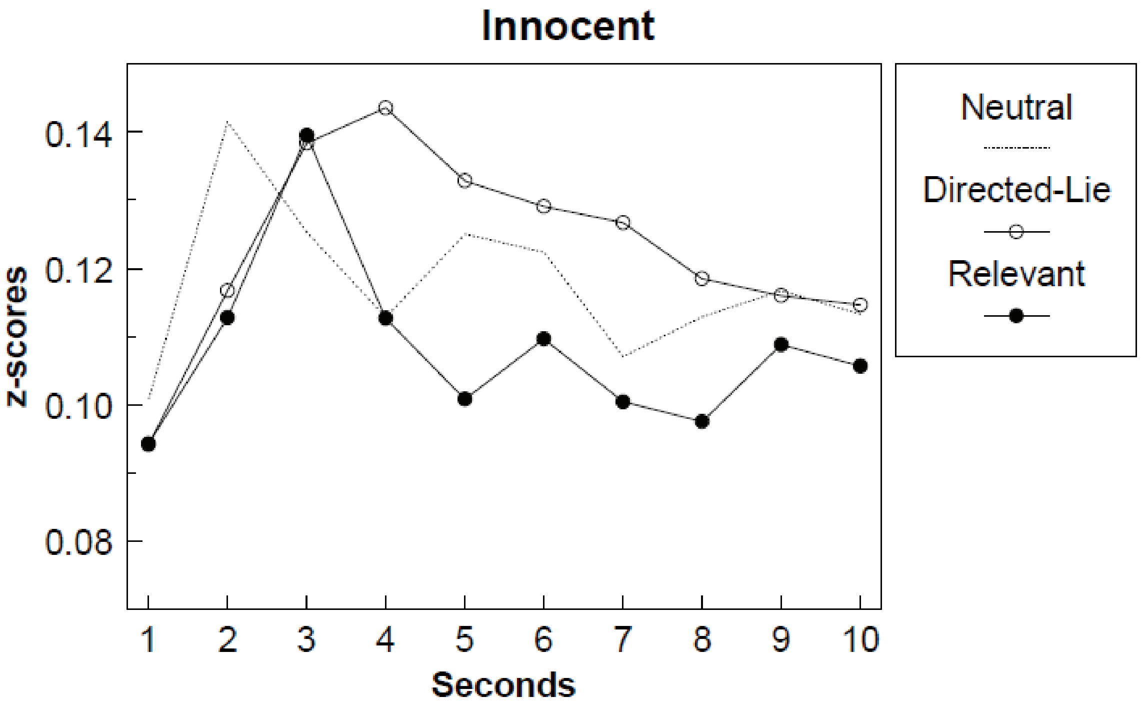


Figure 7b. Respiration excursion associated with neutral, directed-lie, and relevant questions for the innocent (n=30) directed-lie group



Examination of Figure 7a for guilty DL subjects reveals that they responded as expected to the three types of test questions. Neutral questions evoked the greatest variance, or least suppression, in respiratory activity. By comparison, directed-lie questions (open circles) produced some reduction in respiratory activity, and relevant questions (closed circles) produced the greatest suppression in respiratory activity. The response patterns by guilty participants who were given PL tests (not shown) were similar to those obtained from guilty participants given DL tests.

The respiration responses by innocent DL participants to neutral, directed-lie, and relevant questions are shown in Figure 7b. Innocent DL participants did not show the expected pattern of respiration responses to the three types of questions. The responses by innocent DL participants to relevant questions (closed circles) were more suppressed than their responses to directed-lie questions (open circles). Moreover, as compared to neutral questions, directed-lie questions produced an increase in respiratory activity. These data suggest that when innocent subjects are faced with the task of 'responding appropriately' to the directed-lie questions, they tend to breathe more deeply and/or more rapidly than normal.

The data in Figures 7a and 7b indicate that guilty DL participants showed greater respiration activity in response to neutral questions than to directed-lie questions, whereas innocent DL participants showed the opposite pattern. This suggests that the difference between directed-lie and neutral questions might be diagnostic. We explored that possibility by measuring differential respiratory activity to directed-lie and neutral questions rather than using directed-lie and relevant questions. For DL participants, the correlation between the guilt/innocence criterion (guilty coded 0 and innocent coded 1) and the traditional index of differential activity using directed-lie and relevant questions was only .02. However, the correlation between the criterion and differential reactivity using directed-lie and neutral questions was r = .37, *p* < .01. Discrimination between the guilty and innocent DL groups with the new respiration index approached that observed for the PL groups with the traditional respiration index (r = .43). Effects of replacing the traditional respiration index with the new index on computer decisions are described below.

Independent Numerical Evaluations for Probable-Lie and Directed-Lie Tests

Independent numerical evaluations of the first three charts were obtained for participants in the effective feedback conditions. Point-biserial correlations were computed to assess the diagnostic validity of numerical evaluations of respiration, SC, cardiograph, and finger pulse amplitude (McNemar, 1968). The point-biserial correlation (r_{pb}) was obtained between the 3-chart total numerical score and a dichotomous variable that distinguished between guilty (coded 0) and innocent participants (coded 1). The magnitude of the correlation was a measure of the ability of the numerical score to discriminate between the guilty and innocent groups (validity). Coefficient alpha was used to assess the reliability of each physiological measure across the nine comparison/relevant question pairs on the first three polygraph charts. Coefficient alpha was an index of the consistency of numerical scores across the nine comparisons. Validity (r_{pb}) and reliability indices (coefficient alphas) are presented in Table 3 for PL and DL tests.

Table 3. Validity (and internal consistency) of independent numerical evaluations of the first three charts from participants in the effective feedback conditions

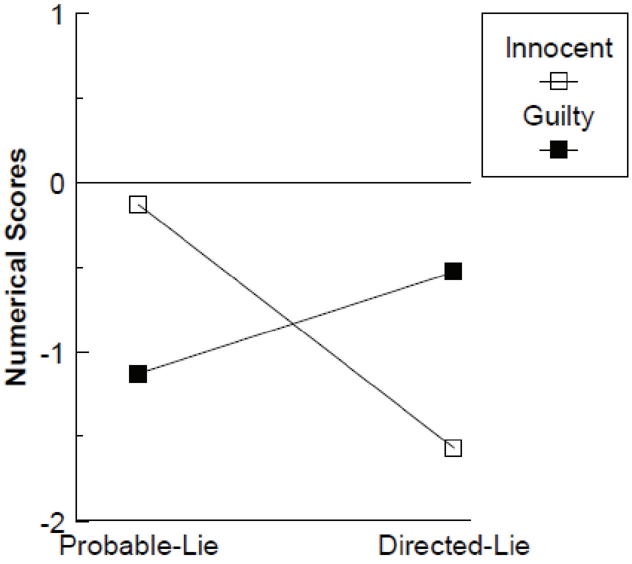
	Probable-lie n = 60	Directed-lie n = 60
Skin Conductance	.73 (.78)	.60 (.74)
Cuff Pressure	.42 (.57)	.31 (.65)
Respiration[a]	.21 (.65)	-.21 (.70)
Finger Pulse Amplitude	.43 (.65)	.44 (.69)

Note: $r_{pb} > .26$ was significant, $p < .05$
[a]The difference between r_{pb} for PL and DL tests was significant, $p < .05$ (McNemar, 1968).

The numerical scores for all components except respiration were significant. In addition, there was a significant difference between the PL and DL tests in the diagnostic validity of numerical scores for respiration. (We recognize the logical inconsistency of concluding, on one hand, that the r_{pb} for PL and DL tests individually do not differ from zero and, on the other hand, that there was a significant difference between the r_{pb} for PL and DL tests. However, r_{pb} for the PL test across all treatment conditions was .27, which was significant at $p < .01$, and the r_{pb} for the DL test across all treatment conditions was -.08, which did not differ from zero. For the entire sample, the difference between the PL and DL tests was still significant.) Mean numerical scores for PL and DL tests are shown in Figure 8.

Participants who received PL tests showed the expected pattern of results. Although one would predict that the mean numerical score for innocent participants would exceed zero, at least the respiration numerical scores were higher for innocent than guilty participants. In contrast, for the DL test, numerical scores were lower for innocent than guilty participants. The findings for the DL test are consistent with the plots of respiration excursion presented in Figures 7a and 7b above. As compared to guilty DL participants, innocent DL participants tended to show more suppression to relevant questions than to DL questions.

Figure 8. Independent numerical evaluations of respiration for probable-lie (n = 60) and directed-lie (n = 60) tests



Discussion

Accuracy of PL and DL Tests. The DL test has several advantages. As compared to PL questions, DL questions are more easily standardized and are less intrusive and embarrassing to the examinee. In addition, prior research suggested that DL tests are at least as accurate as PL tests (Honts & Raskin, 1988; Horowitz et al., 1997).

Comparisons of PL and DL tests in the present study were limited to participants who had received the pretest and effective feedback, since these conditions most closely approximate those in the field. There were no significant differences between the PL and DL tests in distributions of correct, wrong, and inconclusive outcomes for guilty or innocent participants, whether numerical or computer decisions were considered. Excluding inconclusive outcomes, numerical evaluations were 90% correct for the PL test and 84% correct for the DL test; and computer analyses were 90% correct for the PL test and 83% correct for the DL test.

In the Horowitz et al. (1997) study, there were no significant differences between PL and DL tests in decision outcomes, although the results favored the DL test. The decision accuracy obtained for the DL test in the present study was virtually identical to that reported by Horowitz et al.. However, decision accuracy for PL tests was higher in the present study than in the Horowitz study, and the present results tended to favor the PL test. Taken together, the results from the two studies suggest that there is little or no difference between PL and DL tests in their distributions of correct, wrong, and inconclusive outcomes.

Nonparametric comparisons of decision outcomes are not as sensitive to differences between test formats as are parametric analyses of the physiological measurements that underlie those decision outcomes. Planned comparisons of SC, cardiograph, and respiration measures revealed a significant difference between PL and DL tests in the diagnostic validity of the respiration excursion measure. The traditional excursion measure was highly diagnostic for the PL test (r = .43), but it was uncorrelated with the criterion for the DL test (r = -.02). Numerical evaluations of respiration were also more diagnostic for PL tests than for DL tests. These findings are consistent with those reported by Horowitz et al. (1997).

As expected, guilty participants showed greater respiration suppression in response to relevant questions than to comparison questions, and it did not matter if they were given a PL or DL test. In contrast, the respiration responses by innocent participants depended on whether they were given a PL or DL test. The majority of innocent PL participants (52%) showed more respiration suppression in response to PL questions than to relevant questions. In contrast, the majority of innocent DL participants (78%) showed more respiration suppression to relevant questions than to DL questions. Thus, most of the innocent DL participants appeared deceptive on this respiration measure.

Respiration Responses during DL tests. We conducted exploratory analyses of second-by-second changes in respiration excursion associated with neutral, DL, and relevant questions for participants in the effective feedback conditions. Responses to neutral questions served as a baseline of normal respiratory activity. As compared to this baseline, innocent participants breathed more rapidly and/or deeply in response to DL questions. Innocent participants appeared deceptive, not because their respiratory activity in response to relevant questions was suppressed, but because their respiratory activity in response to DL questions increased. Ordinarily, when a person orients to a perceived threat, their breathing slows and becomes shallow (Lynn, 1966). The observed increase in respiratory activity shown by innocent participants in response to DL questions was not the type of orientation response that is typically observed during polygraph examinations (Kircher & Raskin, 1988; Timm, 1982).

During PL tests, innocent subjects are likely to be deceptive to probable-lie questions, and they want to avoid detection on those questions. Innocent subjects do what guilty subjects do when guilty subjects are asked relevant questions; innocent subjects attempt to avoid detection by inhibiting their physiological responses to the probable-lie questions. In contrast, during DL tests, subjects are told that it is important that they appear deceptive on DL questions. The subject's task is different. Rather than trying to avoid detection, subjects try to appear deceptive. Rather than trying to inhibit their reactions to comparison questions, subjects attempt to produce reactions to those questions. To accomplish this task, knowingly or not, innocent subjects alter the one physiological measure over which they have the greatest control, their respiration. PL and DL questions place different demands on subjects, and respiration is sensitive to those differences. When subjects attempt to inhibit responses, their respiration is suppressed, and when they attempt to produce responses, their respiratory activity increases.

Interestingly, innocent subjects appear to use respiration to produce reactions to DL questions, but guilty subjects do not. If respiration suppression is a measure of inhibition, then guilty subjects use the same strategy whether they are given a PL or DL test. In both the PL and DL test, the responses by guilty subjects to relevant questions are more suppressed than their responses to comparison questions, and their responses to comparison questions are more suppressed than their responses to neutral questions. Guilty subjects have but one goal; they want to avoid detection. They attempt to do this by inhibiting their responses to both comparison and relevant questions. Thus, unlike innocent subjects, it does not appear that guilty subjects try to produce reactions to DL questions; they try to inhibit them.

Research by Gross and Levenson (1993) is consistent with the idea that respiration suppression is an indication of deliberate attempts by subjects to suppress the expression of emotional responses. They measured general somatic activity, respiration cycle-time, and respiration amplitude in participants who attempted to inhibit outward displays of negative emotion. Although they found no significant reduction in respiration rate or amplitude, they did find that attempts to suppress emotional displays were associated with reduced somatic activity. The latter finding is consistent with the idea that attempts to inhibit emotional responses are associated with general motor quieting, and this includes a reduction in respiratory activity. Their failure to observe effects on respiration directly may be due to their methods of measuring respiratory activity. Prior research indicates that although measures of respiration rate and amplitude individually have low reliability and validity for the detection of deception, a composite measure of line length from the same data is highly reliable and diagnostic (Kircher & Raskin, 1988; Podlesny & Kircher, 1999).

The findings from the present study and those from Horowitz et al. (1997) argue that the procedures for evaluating respiration responses recorded during DL tests should be modified to avoid false positive errors. Overall, numerical evaluations of respiration responses during DL tests were not diagnostic, and in the standard effective-feedback condition, numerical scores for respiration were almost significantly more negative for innocent subjects than for guilty subjects. Obviously, it would be better to drop the evaluations of respiration responses altogether than to include them if they work against a valid decision.

The present study examined an alternative measure of differential respiration reactivity for DL tests. Guilty and innocent subjects could not be distinguished based on their respiration responses to DL and relevant questions (r = -.02). However, second-by-second plots of respiration excursion revealed that innocent subjects showed more respiratory activity in response to DL questions than to neutral questions, whereas guilty subjects showed more respiratory activity in response to neutral questions than DL questions. The difference between DL and neutral questions was diagnostic. It correlated .37 with the criterion and made significant contributions to several proposed statistical models for DL tests (described below).

Initial results with the new respiration index for DL tests are promising. However, the decision to measure the difference between DL and neutral questions was made after viewing the second-by-second plots of respiration excursion (post hoc). Although the measure correlated with the criterion in the present study, its diagnostic validity should be established in an independent sample of cases. Moreover, to our knowledge, there is no precedent for measuring the difference between comparison and neutral questions. Polygraph tests are used to draw inferences about the veracity of subjects' answers to relevant questions. Use of the difference between reactions to two types of questions, neither of which pertain to the matter under investigation, is indirect. Although a computer easily could be programmed to measure and use responses to neutral questions, numerical evaluators would find it more difficult to evaluate responses not only to DL and relevant questions but also to neutral questions. Finally, if responses to neutral questions were to be used for diagnosis, the question sequence would have to contain enough neutral questions to ensure that the reliability of measures of responses to those questions is adequate.

Study 3: New Physiological Measures and Computer Models for the Detection of Deception

Computer indices of differential reactivity were obtained from the PL and DL participants in the standard effective-feedback condition. Features were extracted from traditional measures (SC, cardiograph, and finger pulse amplitude) as well as several new measures (skin potential, systolic blood pressure, diastolic blood pressure, mean blood pressure, and vagal tone).

Skin potential responses are often biphasic. An initial increase in negativity is followed by a rapid decrease in negativity that drops well below the initial baseline and then slowly recovers to baseline. Features were extracted from the original skin potential waveform, which represented increases in negativity. The skin potential waveform was then reflected (multiplied by –1), and a second set of features was extracted from the reflected waveform. The reflected waveform represented increases in positivity.

Vagal tone was obtained for eight 10-second intervals. The beginning of the 10-second measurement interval was varied in 2-second increments from 0 to 14 seconds following stimulus onset. Vagal tone was measured using the Porges et al. (1980) algorithm.

Reliability and Validity of Physiological Measures

Point-biserial correlations were computed to assess the diagnostic validity of each physiological measure, and coefficient alpha was used to assess the reliability of each physiological measure across the nine comparison/relevant question pairs on the first three polygraph charts. The validity (r_{pb}) and reliability indices (coefficient alphas) are presented in Table 4 for PL tests and in Table 5 for DL tests.

Table 4. Validity (and reliability) of differential reactivity indices for PL tests (n = 60)

Feature	Skin Conductance	Cardiograph	Finger Pulse Amplitude	Skin Potential Negative	Skin Potential Positive	Systolic Blood Pressure	Diastolic Blood Pressure	Mean Blood Pressure	Vagal Tone	**Interval (sec)**
Peak Amplitude	**.73 (.80)**	**.45 (.56)**	.43 (.54)	.30 (.50)	.60 (.80)	.45 (.55)	.46 (.57)	.48 (.57)	.00 (.38)	**0-10**
Area	.71 (.78)	.48 (.51)	.39 (.66)	.04 (.63)	.48 (.78)	.29 (.55)	.28 (.61)	.28 (.57)	-.13 (.31)	**2-12**
Latency	-.48 (.50)	-.28 (.32)	-.23 (.22)	.29 (.54)	-.16 (.40)	-.24 (.39)	.-29 (.38)	-.24 (.42)	-.12 (.41)	**4-14**
Rise Time	.32 (.64)	.29 (.42)	.14 (.54)	-.27 (.71)	.14 (.66)	.08 (.10)	-.01 (.33)	-.03 (.09)	-.01 (.39)	**6-16**
Recovery Time	.63 (.66)	.21 (.24)	.08 (.47)	-.05 (.59)	-.03 (.61)	.18 (.47)	.18 (.39)	.21 (.31)	.02 (.27)	**8-18**
Rise Rate	.67 (.73)	.10 (.25)	.11 (.54)	.50 (.62)	.64 (.61)	.22 (.34)	.27 (.40)	.37 (.21)	.20 (.21)	**10-20**
Recovery Rate	.67 (.66)	.23 (.31)	.14 (.27)	.05 (.41)	.47 (.48)	-.10 (.32)	.08 (.24)	.14 (.37)	.22 (.20)	**12-22**
Excursion	.75 (.77)	.30 (.40)	.20 (.50)	.68 (.75)	.68 (.75)	.30 (.56)	.44 (.56)	.41 (.56)	.15 (.33)	**14-24**
Variance	.72 (.80)	.34 (.41)	.39 (.51)	.71 (.73)	.71 (.75)	.47 (.56)	.56 (.58)	.51 (.57)		

Note: -.26 <r_{pb} <.26 were not significant at p < .05. Variables used by CPS to make decisions are highlighted in bold.

Table 5. Validity (and reliability) of differential reactivity indices for DL tests (n = 60)

Feature	Skin Conductance	Cardiograph	Finger Pulse Amplitude	Skin Potential Negative	Skin Potential Positive	Systolic Blood Pressure	Diastolic Blood Pressure	Mean Blood Pressure	Vagal Tone	**Interval (sec)**
Peak Amplitude	**.63 (.80)**	**.36 (.60)**	.53 (.59)	.37 (.67)	.61 (.72)	.41 (.52)	.44 (.55)	.43 (.56)	.07 (.08)	**0-10**
Area	.55 (.80)	.34 (.62)	.53 (.70)	.20 (.65)	.52 (.67)	.34 (.52)	.45 (.46)	.38 (.51)	.09 (.35)	**2-12**
Latency	-.42 (.34)	-.36 (.38)	-.21 (.01)	.10 (.11)	.05 (.19)	-.24 (.32)	-.14 (.00)	-.18 (.25)	.02 (.40)	**4-14**
Rise Time	.42 (.59)	.40 (.48)	.31 (.46)	-.25 (.53)	.22 (.52)	.04 (.13)	-.25 (.33)	-.06 (.26)	.15 (.32)	**6-16**
Recovery Time	.59 (.61)	.07 (.31)	.17 (.45)	-.04 (.52)	.32 (.50)	.24 (.30)	.39 (.39)	.27 (.46)	.15 (.48)	**8-18**
Rise Rate	.65 (.76)	-.05 (.18)	.03 (.49)	.50 (.59)	.69 (.46)	.20 (.24)	.46 (.56)	.30 (.49)	.21 (.48)	**10-20**
Recovery Rate	.47 (.70)	.34 (.38)	.08 (.27)	.00 (.32)	.44 (.28)	.10 (.20)	.18 (.38)	.23 (.45)	.17 (.50)	**12-22**
Excursion	.59 (.79)	.21 (.54)	.01 (.35)	.68 (.79)	.68 (.79)	.38 (.52)	.47 (.57)	.48 (.56)	.18 (.48)	**14-24**
Variance	.59 (.82)	.20 (.32)	.26 (.27)	.63 (.79)	.63 (.79)	.45 (.36)	.53 (.46)	.58 (.35)		

Note: -.26 <r_{pb} <.26 were not significant at p < .05. Variables used by CPS to make decisions are highlighted in bold.

For the PL test, the point-biserial correlation and coefficient alpha for respiration excursion were .43 and .73, respectively. For the DL test, point-biserial correlation and coefficient alpha for respiration excursion were -.02 and .60.

As expected, skin potential excursion was among the most promising of the new measures. For the PL test, skin potential excursion was almost as highly correlated with the criterion (r_{pb} = .68) as was SC amplitude (r_{pb} = .73). For the DL test, it was more highly correlated with the criterion (r_{pb} = .68) than was SC amplitude (r_{pb} = .63). Overall, changes in the positivity of the skin potential signal appeared to be more useful for detecting deception than were changes in negativity. To measure changes in positivity, the original skin potential waveform had been multiplied by –1. As expected and may be seen in Table 7 and Table 8, this transformation had no effect on the excursion and variance measures.

Consistent with findings by Podlesny and Kircher (1999), features extracted from the blood pressure recordings tended to be more diagnostic than those obtained from the cardiograph. Whereas Podlesny and Kircher (1999) found that increases in systolic blood pressure were significantly more diagnostic than were increases in the baseline of the cardiograph, the differences observed in the present study were not significant. None of the vagal tone measures was correlated significantly with the criterion for either the PL or DL test.

Contributions of New Physiological Measures to Computer Decision Models

The current computer model combines scores on SC amplitude, cardiograph amplitude, and respiration excursion measures by means of a discriminant function. For DL tests, we explored the possibility of replacing the traditional respiration excursion measure with one that compares responses to directed-lie and neutral questions. We also explored the possibility of adding new measures to computer decision models. Finally, we compared decision outcomes produced by discriminant functions and logistic regression models that contained the same variables.

Effects of New Respiration Index on Computer Decisions. As noted above, participants breathe differently during DL and PL tests. Consequently, the traditional approach that compares respiration responses to comparison and relevant questions does not work for DL tests. However, differential reactivity to directed-lie and neutral questions was correlated with the criterion and might be used in place of the traditional index. We created a decision model for the DL effective-feedback group using the standard SC amplitude, cardiograph amplitude, and traditional respiration excursion measures. We then replaced the traditional respiration index with the new respiration measure and developed a second decision model. Outcomes using the traditional and new respiration measures are presented in left half of Table 6.

Logistic regression is an alternative to discriminant analysis. Logistic regression is similar to discriminant analysis in that it weighs physiological measures and combines them into a single score that is optimal for separating groups of known truthful and deceptive individuals. However, discriminant analysis and logistic regression use different statistical methods for deriving variable weights, and the assumptions that underlie the use of logistic regression are less restrictive than those that underlie discriminant analysis. Since logistic regression also provides probabilities of truth (or deception), participants could be classified according to the same rules developed for the discriminant function models. Table 6 reports the outcomes obtained from application of those decision rules. Specifically, the participant was classified as truthful if the probability of truthfulness after three charts exceeded .70. The participant was classified as deceptive if the probability of truthfulness after three charts was less than .30. If the probability fell between those two cutoffs, a new probability was computed using all five polygraph charts. If the probability of truthfulness based on five charts exceeded a .70 or .30 cutoff, the individual was classified as truthful or deceptive. Otherwise, the test was considered inconclusive.

Overall decision accuracy increased when the new respiration index was substituted for the old index. The logistic regression model showed the greatest improvement; decision accuracy increased from 81% to 88% for innocent participants and from 86% to 89% for guilty participants. Although use of the new respiration index had a modest effect on decision accuracy, its contributions to both discriminant functions (3-chart t(56) = 2.12, $p < .05$ and 5-chart t(56) = 2.32, $p < .05$) and both logistic regression models (3-chart Wald statistic = 5.06, $p < .05$ and 5-chart Wald statistic = 5.77, $p < .05$) were statistically significant.

Table 6. Percent outcomes for DL tests using discriminant functions or logistic regressions that contained the traditional or new index of respiration activity

		Discriminant Functions					**Logistic Regressions**			
		N	**C**	**W**	**I**	**CD**	**C**	**W**	**I**	**CD**
Traditional	Innocent	30	73	17	10	81	70	17	13	81
	Guilty	30	73	10	17	88	63	10	27	86
New	Innocent	30	70	17	13	81	70	10	20	88
	Guilty	30	83	7	10	93	80	10	10	89

Contributions of New Physiological Measures to Decision Models. SC latency, FPA amplitude, FPA area, and skin potential excursion were selected for additional analysis. Each of these variables was added to a decision model that included our three standard measures: SC amplitude, cardiograph amplitude, and respiration excursion. SC latency was chosen because it was significantly correlated with the criterion for both PL and DL tests, and it was relatively independent of SC amplitude, which was already in the decision model. Use of SC latency also would not require the collection of any new channels of physiological activity. FPA amplitude and FPA area were considered because they were highly correlated with the criterion for both PL and DL tests, and some polygraph examiners already collect and score FPA data. FPA amplitude was more highly correlated with the criterion than FPA area, but FPA area was more reliable. Skin potential excursion was selected because it too was highly correlated with the criterion for both PL and DL tests and was highly reliable. No blood pressure measures were selected because they are highly correlated with cardiograph measures (Podlesny & Kircher, 1999), and the technology for recording blood pressure continuously is expensive and not readily available on commercial polygraphs.

For PL tests, a discriminant function was created using the SC amplitude, cardiograph amplitude, and the original respiration excursion measures from the first three charts. This standard set of measures was used for DL tests, except that the new respiration index that compared reactions to directed-lie and neutral questions was substituted for the original respiration index. The proportion of criterion variance explained by the SC, cardiograph, and respiration measures for the PL and DL tests were .633 and .447, respectively. SC amplitude, cardiograph amplitude, and the original or the new respiration measures were also used to create logistic regression models for PL and DL tests. The proportions of variance explained by the standard measures in the logistic regression models for the PL and DL tests were .591 and .432. Table 7 reports the increment in proportion of criterion variance explained (ΔR^2) when each of the selected physiological measures was then combined with the three standard measures in the discriminant function or logistic regression model.

For PL tests, none of the new variables contributed significantly to a computer model that contained SC amplitude, cardiograph amplitude, and respiration excursion. In contrast, for DL tests, several of the new measures contributed significantly to the computer models. For example, skin potential excursion increased the proportion of variance explained by the logistic regression model from .432 to .489 (5.7%).

Table 7. Increments in proportion of criterion variance explained (ΔR²) by new measures

	Probable-Lie Tests		Directed-Lie Tests	
	Discriminant Function ΔR^2	Logistic Regression ΔR^2	Discriminant Function ΔR^2	Logistic Regression ΔR^2
SC latency	.010	.007	.010	.016
FPA amplitude	.003	.001	.045*	.054*
FPA area	.001	.000	.038	.046*
Skin potential excursion	.001	.000	.048*	.057*

*p < .05

The effects of adding FPA amplitude or skin potential excursion to the computer model on outcomes of DL tests are presented in Table 8. The first two rows of outcomes are for a decision model that used discriminant functions or logistic regression models composed of the standard three physiological measures (SC amplitude, cardiograph amplitude, and the new respiration excursion measure). The next two rows show the outcomes for a model that adds FPA amplitude to the standard three measures. The last two rows of the Table 8 show the outcomes for a model that adds skin potential excursion to the standard three measures.

On average, the addition of FPA amplitude or skin potential excursion to the decision models for DL tests increased the percentage of correct outcomes by 13% and reduced the percentage of inconclusive outcomes by 7%. The mean inconclusive rate for the logistic regression models (M = 10%) was higher than that for the discriminant

Table 8. Percent Correct (C), Wrong (W), Inconclusive (I), and Correct Decisions excluding inconclusives (CD) for new directed-lie computer models (N = 60)

		Discriminant Functions				**Logistic Regression Models**			
		C	**W**	**I**	**CD**	**C**	**W**	**I**	**CD**
SC amplitude Cardiograph amplitude Respiration excursion	Innocent	70	17	13	81	70	10	20	88
	Guilty	80	10	10	89	80	10	10	89
SC amplitude Cardiograph amplitude Respiration excursion FPA amplitude	Innocent	77	17	7	82	80	13	7	86
	Guilty	87	10	3	90	80	7	13	92
SC amplitude Cardiograph amplitude Respiration excursion Skin potential excursion	Innocent	80	13	7	86	83	13	3	86
	Guilty	90	7	3	93	87	7	7	93

functions (M = 7%). Excluding inconclusive outcomes, decision accuracy for the logistic regression models (M = 89%) was slightly higher than that produced by the discriminant functions (M = 87%). However, the differences between the two approaches were small and non-significant.

Discussion

New Physiological Measures. Study 3 assessed the psychometric properties of vagal tone, skin potential, and blood pressure for the detection of deception. Vagal tone was assessed using the Porges-Bohrer algorithm (Porges et al., 1980) for a series of partially overlapping 10-second intervals. The beginning of the measurement interval was systematically varied in 2-second increments from 0 to 14 seconds relative to stimulus onset. The first 10-second interval began at stimulus onset and ended at the 10th post-onset second. The next 10-second interval began at the 2nd post-onset second and ended at the 12th post-onset second, and so on.

For both PL and DL tests, the diagnostic validity of the vagal tone measure reached a maximum correlation of about .20 for the interval that began at about the 10th post-onset second. Since the vagal tone results were obtained separately for PL and DL tests and analyses were limited to the effective-feedback conditions, the sample sizes were not sufficient to achieve statistical significance (n = 60 per group). Had the samples been combined, the sample size would have been 120, and the maximum observed correlations would have been statistically significant. Moreover, the vagal tone results for independent samples of PL and DL participants were similar, which suggests that the effects on vagal tone were small but reliable.

It is interesting to note that the vagal tone measure was most diagnostic in the latter part of the post-onset recording interval. These data are consistent with the idea that parasympathetic effects on the heart were greatest when the subject was recovering from a strong sympathetic response to a test question. If the vagal effects were secondary to strong sympathetic response, then it is unclear if vagal tone would contribute to a decision model composed primarily of measures of sympathetic activity. The possibility was not evaluated since vagal tone was not sufficiently correlated with the criterion to be considered for inclusion in a new decision model. To be considered, a new variable had to be correlated significantly with the criterion.

Skin potential was also evaluated in the present study. Consistent with prior research (Raskin & Kircher, 1990), skin potential was highly diagnostic of truth and deception. For PL tests, skin potential excursion (line length) correlated .68 with the criterion and was almost as highly correlated with the criterion as was SC amplitude (r = .73). For DL tests, skin potential excursion yielded the greatest correlation with the criterion (r = .68). Although skin potential was highly correlated with SC amplitude (M_r = .85), it made significant contributions to new decision models for DL tests that included SC amplitude. In one new model, skin potential increased the accuracy of definite decisions by 4.5% and reduced inconclusive outcomes by 6.5%. In another new decision model, skin potential had little effect on the accuracy of definite decisions but reduced inconclusive outcomes from 15% to 5%.

Findings from two laboratory studies now indicate that skin potential is a promising new measure for the detection of deception (cf., Raskin & Kircher, 1990). The usefulness of skin potential for field polygraph examinations will not be established unless field polygraph instruments are outfitted with a circuit for recording skin potential and computer software is modified to collect and store it. Fortunately, skin potential can be measured with a simple voltmeter. The circuit is completely inactive and simpler than the skin conductance or resistance circuits that are currently used on field polygraphs. A disadvantage in recording skin potential is the need for wet electrodes. Another disadvantage is the method for measuring the skin potential response. Since the response is biphasic, simple measurements of negative or positive wave amplitude are not as effective as the measurement of line length (excursion). Although at least one numerical scoring system has been developed that requires human evaluators to measure line length with a planimeter, it has not been recommended for general use (Dutton, 2000; Krapohl & Norris, 2000). Use of a computer would facilitate the measurement of skin potential responses.

Blood pressure was recorded continuously from a finger using a Finapres arterial pressure monitor. This was the second of two laboratory studies to investigate the validity of changes in blood pressure for comparison-question polygraph tests. Previously, Podlesny and Kircher (1999) conducted PL tests and found that increases in systolic blood pressure were significantly more diagnostic than were increases in the baseline of the cardiograph. For PL participants in the present study, the validity coefficients for the Finapres and cardiograph were statistically indistinguishable. For the DL test, the results tended to favor the Finapres, but again, the differences were not statistically significant. Taken together, the findings from the these two laboratory studies suggest that we could replace the cardiograph with the Finapres or a similar device and expect a small improvement in our ability to discriminate between truthful and deceptive subjects. More importantly, since the Finapres can be operated for long periods of time without discomfort, it would allow for the presentation of more test questions and might increase the reliability of the physiological measures. It would offer greater flexibility in the construction of polygraph tests that might lead to more valid polygraph outcomes. For instance, polygraph examiners currently stop collecting data every 5 or 6 minutes to release the pressure in the cardiograph cuff. The cuff is deflated to avoid vasocongestion in the arm below the cuff and pain. If the cardiograph were not used, an extended uninterrupted series of questions could be presented that might reveal diagnostic differences between truthful and deceptive subjects in the habituation rates of their physiological responses to comparison and relevant questions (Ben Shakhar, Lieblich, & Kugelmass, 1975; Kircher, Raskin, & Honts, 1984).

New Decision Models. New decision models were developed for PL and DL tests that used either discriminant functions or logistic regression models to compute the probability of truthfulness from physiological measures. A standard model composed of SC amplitude, cardiograph amplitude, and respiration excursion was developed using the participants in the effective feedback conditions. Attempts were then made to determine if any of four new measures would improve the accuracy of polygraph outcomes when added to the standard model. The four new measures included the latency from stimulus onset to the onset of the SC response, the amplitude of the FPA response, the area under the FPA response, and skin potential excursion.

For the PL test, the standard three physiological measures accounted for about 60% of the variance in the criterion, and none of the new measures added to the proportion of variance explained by either the discriminant function or logistic regression model. For the DL test, the standard three measures accounted for about 44% of the variance in the criterion, and several of the new measures made significant contributions to the discriminant and logistic regression models. Individually, FPA amplitude and skin potential excursion increased the proportion of variance explained by the discriminant function; and FPA amplitude, FPA area, and skin potential increased the proportion of variance explained by the logistic regression model. The most diagnostic model for the DL test consisted of SC amplitude, cardiograph amplitude, respiration excursion, and skin potential excursion. The addition of skin potential excursion increased the accuracy of definite decisions from 81% to 86% for innocent participants and from 89% to 93% for guilty participants, and reduced inconclusive outcomes from 12% to 5%. The accuracy of outcomes for the new DL model was comparable to that obtained with the original CPS discriminant function for the PL test.

Overall, there was little difference between decision models based on discriminant functions and logistic regression models. The discriminant functions typically accounted for slightly more variance in the criterion, but the logistic regression models tended to achieve a slightly better balance of false positive and false negative errors. These findings are consistent with those reported previously (Kircher, Raskin, Honts, & Horowitz, 1994). Kircher et al. found that when the same variables were used in a discriminant function and a logistic regression model, the correlation between the probabilities of truthfulness produced by the two models approached 1.0. In contrast, Devitt and Honts (1993) analyzed many of the same data and concluded that a logistic regression model outperformed a discriminant function. However, Devitt and Honts did not indicate if they used the same physiological measures to develop their logistic and discriminant function models. Therefore, it is unclear if the observed difference in decision accuracy was due to the method of analysis (discriminant analysis versus logistic regression) or to differences in the physiological variables used by the two models.

Study 4: Effects of Personality on Polygraph Outcomes

Participants were asked to rate the accuracy of polygraph tests on two occasions. They gave one accuracy rating prior to learning if they were in the guilty or innocent treatment condition and gave another rating after their polygraph test but before they were informed of the decision. The rating scale ranged from 1 (Not At All Accurate) to 9 (Perfectly Accurate). Participants rated the accuracy of polygraph tests for deceptive examinees in general, truthful examinees in general, and for themselves in particular.

Effects of Feedback on Ratings of Polygraph Accuracy. Feedback was expected to alter participants' ratings of polygraph accuracy. Specifically, we expected effective feedback to increase ratings of polygraph accuracy and ineffective feedback to decrease ratings of polygraph accuracy.

A split-plot ANOVA was conducted with Occasions as the repeated measure (pretest and posttest) and Guilt, Feedback, and Test Type as between-group factors. The expected Feedback X Occasion interactions did not materialize, but the means were generally in the expected direction and approached significance for ratings concerning truthful examinees, F(1, 160) = 2.35, $p < .13$ and for ratings concerning the participant, F(1, 160) = 2.97, $p < .09$.

The ANOVA revealed an effect of Occasions on participant ratings of polygraph accuracy for themselves. Participants' mean rating of polygraph accuracy for themselves increased from 6.15 at pretest to 6.45 at posttest, F(1, 320) = 10.3, p < .01. There was also an Occasion X Test Type interaction effect on ratings of polygraph accuracy for guilty people in general, F(1, 320) = 5.56, p < .02. Mean ratings by PL participants increased from 6.41 to 6.62, whereas ratings by DL participants decreased from 6.42 to 6.16. After having taken a PL test, participants' ratings of polygraph accuracy on guilty people in general increased slightly, but after having taken a DL test, participants' ratings of accuracy on guilty people decreased.

Other Analyses of Ratings of Polygraph Accuracy. Correlations were obtained between demographic measures (Age, Sex, Ethnicity, Education) and pretest ratings of polygraph accuracy as well as pre-post change scores. Several correlations were significant but demographic measures never accounted for more than 2% of the variance in participants' ratings of polygraph accuracy or change scores.

Interactions of Feedback and Personality. People who wish to make a good impression, are more trusting, or are more anxious may be more or less affected by the nature of the feedback provided by the polygraph examiner. Additional analyses were conducted to test if differences between effective and ineffective groups were related to personality characteristics. A separate multiple regression analysis was performed to test for a three-way interaction effect of Guilt X Feedback X Personality Dimension on each of three physiological measures. To illustrate, Figure 9 shows a hypothetical three-way interaction of Guilt, Feedback, and Interpersonal Trust. It is consistent with the idea that guilty and innocent subjects who are more trusting of others will be more convinced by the effective feedback typically provided by the polygraph examiner. If effective feedback is important for the detection of deception, then decision accuracy should improve as people are more convinced by the feedback. The left panel of Figure 9 shows that when subjects are given effective feedback, discrimination between guilty and innocent subjects improves as the level of trust increases. Conversely, the right panel shows that when subjects are given feedback that the test is ineffective, discrimination declines as the level of trust increases.

The three dependent physiological measures were indices of differential reactivity for SC amplitude, cardiograph amplitude, and respiration excursion. The independent variables consisted of three main effects (Guilt, Feedback, Personality Characteristic), the three two-way interactions, and the three-way interaction of interest. Guilt and Feedback were effect-coded (+1, -1), the personality measure was transformed to a set of *z*-scores (centered), and cross-products provided the two-way and three-way interaction terms.

Figure 9. Scatterplots showing a hypothetical Guilt X Feedback X Trust interaction

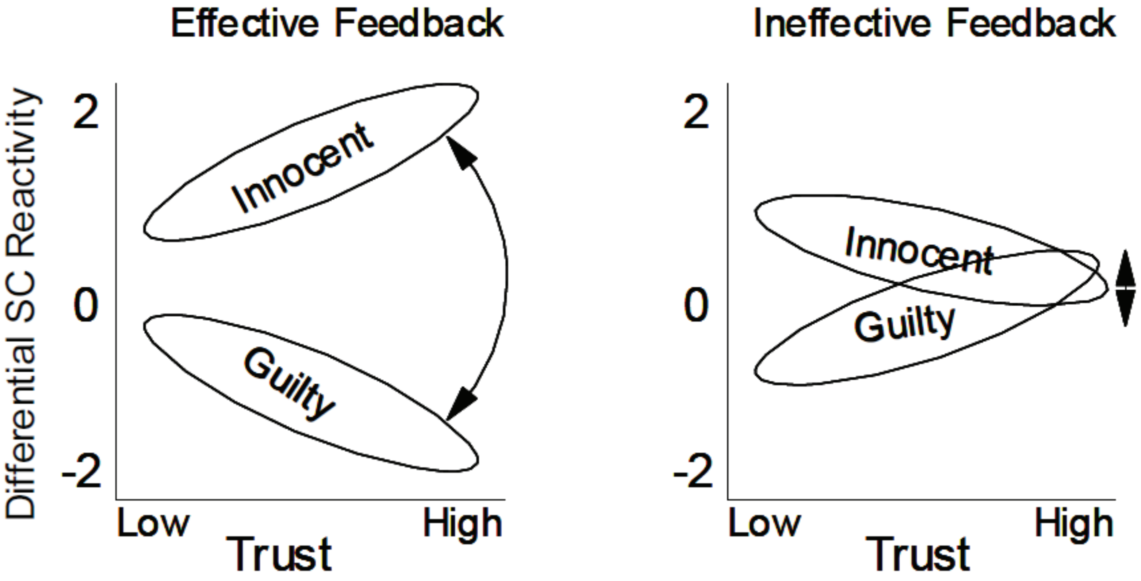


Table 9 presents the proportion of variance in the physiological measure explained by the main and two-way interaction terms (R^2) and the increment in proportion of variance explained when the three-way interaction term was added to the equation (Δ R^2). The 84 PL participants were those who received effective feedback (n = 60) plus those who received ineffective feedback (n = 24). Similarly, the 84 DL participants were those who received effective (n = 60) and ineffective feedback (n = 24).

Only one of the 24 increments in proportions of variance explained by the three-way interaction (Δ R^2) was statistically significant. The effect was small, and it did not generalize to the DL test or to other physiological measures.

Discussion

Study 4 explored the possibility that the effects of the feedback vary depending on the individual's level of trust, anxiety, or need for social approval. With the possible exception of social desirability, there was little evidence that the effects of positive and negative feedback on physiological responses are moderated by individual differences on these dimensions. Since only one of 24 tests for moderation effects was significant, additional research should be conducted to confirm that the effect is reliable. The present results add to a growing literature that suggests that if the accuracy of polygraph outcomes depends on dimensions of personality, the effects are likely to be small (e.g., Bradley & Janisse, 1981; Honts et al., 1985; Patrick & Iacono, 1989; Raskin & Hare, 1978).

Table 9. Proportions of variance in physiological measures explained by main and two-way interactions of Guilt, Feedback, and Personality Dimension (R^2) and increments in variance explained by the three-way Guilt X Feedback X Personality Dimension interaction (Δ R^2)

		Probable-lie (n = 84)		*Directed Lie (n = 84)*	
Personality Dimension	**Dependent Variable**	**R^2**	**ΔR^2**	**R^2**	**ΔR^2**
Social Desirability	SC Amplitude Cardio Amplitude Respiration Excursion	.541** .172** .200**	.000 .053* .000	.469** .206** .041	.004 .005 .000
Interpersonal Trust	SC Amplitude Cardio Amplitude Respiration Excursion	.555** .258** .195**	.001 .000 .002	.449** .221** .067	.001 .022 .019
Trait Anxiety	SC Amplitude Cardio Amplitude Respiration Excursion	.535** .175* .188*	.007 .016 .000	.431** .198** .048	.007 .000 .004
State Anxiety	SC Amplitude Cardio Amplitude Respiration Excursion	.549** .215** .185*	.004 .016 .007	.431** .207** .030	.001 .002 .001

**p < .01
* p < .05

General Limitations

As always, it is important to note that the data for the present study were obtained from subjects who participated in a mock crime experiment. Whether these laboratory results are representative of the field is an open question. For PL tests, we previously compared data from our lab and field polygraph studies and found that differences between probable-lie and relevant questions in the field sample were generally shifted in the negative direction (Kircher, Raskin, Honts, & Horowitz, 1994). The truthful and deceptive field suspects appeared more deceptive on their polygraph tests than did the truthful and deceptive laboratory subjects. Although the differences between comparison and relevant questions were more negative, the separation between truthful and deceptive individuals was similar for the lab and field samples. The various indices of differential reactivity were as diagnostic for the field sample as they were for the lab sample. In addition, the variances and covariances among various indices of differential reactivity in the lab and field samples were indistinguishable. These findings suggest that the results of the present study can be used to make reasonable inferences about the effects of various pretest procedures on the outcomes of probable-lie polygraph examinations in the field.

For DL tests, the generalizability of results from mock crime experiments is less certain. To our knowledge, there have been no systematic comparisons of the mean and covariance structures of physiological measures collected in laboratory and field settings. To that end, efforts should be made to archive computer files of DL tests. Over time, a sample of verified criminal cases could be developed that would allow for tests of the generalizability of polygraph results across settings.

Sixteen percent of individuals assigned to the guilty condition refused to participate after receiving instructions to commit the mock theft, whereas none of the individuals assigned to the innocent condition refused to participate. The refusal of individuals assigned to the guilty condition was probably an indication of the perceived difficulty of task.

Participants were alone in an unfamiliar environment. They received tape-recorded instructions to wait for a secretary to leave her office and steal $20 from her purse. The taped instructions indicated that they should not tell anyone that they were participating in an experiment and that they should prepare an alibi in case they were caught. They had no face-to-face contact with anyone who could assure them that their participation would not result in real consequences. Attrition from the guilty condition was an indication that our efforts to achieve a high level of personal involvement and some level of realism were successful. Although differential attrition from one experimental group may introduce a selection artifact and is a threat to construct validity (Cook & Campbell, 1976), we observed no significant differences between guilty and innocent participants on any demographic or personality measures.

Summary and Recommendations

The present study tested if a preliminary demonstration test and feedback that the polygraph is effective improves the accuracy of polygraph outcomes. The pretest was beneficial. As compared to a no-pretest control condition, the demonstration test and feedback increased the accuracy of decisions from 77% to 90% for the PL test and from 75% to 83% for the DL test. To maximize decision accuracy, polygraph examiners should continue to administer the pretest.

The results of the present study also suggest that for PL tests, administration of the pretest is sufficient; it is not necessary to try to convince subjects that they showed their strongest reactions when they lied during the pretest. If it is not necessary to convince the examinee that the polygraph is effective prior to conducting a PL test, then the practice may be discontinued. There may be differences among polygraph examiners in their ability to convince examinees that the examinees' deception was revealed during the pretest, and these differences might affect the accuracy of outcomes from polygraph examinations. If polygraph examiners are not expected to convince the examinee of the effectiveness of the polygraph prior to the test, a potential source of variance in the protocol can be eliminated.

We are unsure about some of the results from DL tests. As noted above, the pretest coupled with effective feedback improved decision accuracy by about 8%. For the DL test, we recommend that polygraph examiners continue to conduct the pretest and then to inform the examinee that the polygraph clearly indicated when they were truthful and deceptive.

There were no significant differences between PL and DL tests in the accuracy of polygraph outcomes. However, the respiration responses of innocent subjects to directed-lie questions were unlike those of innocent subjects to probable-lie questions. When innocent subjects were presented with directed-lie questions, their respiratory activity increased. These findings suggest that numerical scoring rules and computer algorithms for probable-lie tests are not optimal for DL tests. Specifically, subjects should not be considered deceptive if they show greater suppression in response to relevant questions than to directed-lie questions. In light of these findings, we recommend that respiration responses not be numerically evaluated or computer scored until it can be established with verified field polygraph examinations that those methods are appropriate for DL tests.

The present study replicated prior research indicating that changes in skin potential and arterial blood pressure in the finger are highly diagnostic of truth and deception. We recommend that efforts be made to collect skin potential data from field suspects. We also recommend that a technology be developed for measuring blood pressure during field polygraph examinations that can be used instead of the cardiograph.

Finally, systematic efforts should be made to develop a national database of field polygraph examinations. Questions concerning the generalizability of laboratory results could be addressed with the accumulation of confirmed DL and PL examinations. The development of such a database also would facilitate additional research and development of polygraph techniques. For example, with a sufficient number of confirmed field cases, it would be possible to compare different types of comparison questions, test formats, subject characteristics, crimes, and computer algorithms for detecting deception. Such a database would not only facilitate research on best practices but also contribute to our understanding of factors that affect the accuracy of polygraph tests and contribute to the development of a well-articulated theory of detection of deception.

References

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

Barland, G. H., Honts, C. R. & Barger, S. D. (1989). Studies of the accuracy of security screening polygraph examinations. Research Division Department of Defense Polygraph Institute, Fort McClellan, AL.

Bell, B. G., Raskin, D. C., Honts, C. R., Kircher J. C. (1999). The Utah numerical scoring system. *Polygraph*, 28, 1-9.

Ben Shakhar, G., Lieblich, I., and Kugelmass, S. (1975). Detection of information and GSR habituation: An attempt to derive detection efficiency from two habituation curves, *Psychophysiology*, 12, 283-288.

BMDP Statistical Software Manual (1981). W. J. Dixon (Ed.) Berkeley, CA: University of California Press.

Bradley, M. T. & Janisse, M. P. (1981). Accuracy demonstrations, threat, and the detection of deception.: Cardiovascular, electrodermal, and pupillary measures. *Psychophysiology*, 18, 307-314.

Crowne, D. P. & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.

Davis, R. C. (1961). Physiological responses as a means of evaluating information. In A. D. Biderman & H. Simmer (Eds.), *The manipulation of human behavior* (pp.142-168). New York: Wiley.

Dawson, M. E. (1980). Physiological detection of deception: Measurement of responses to questions and answers during countermeasure maneuvers. *Psychophysiology*, 17, 8-17.

Devitt, M. K. & Honts, C. R. (1993). Multivariate classifiers perform as well as experts in the detection of deception. Paper presented at the Fifth Annual Convention of the American Psychological Society, Chicago, June.

Dutton, D. W. (2000). Guide for performing the objective scoring system. *Polygraph*, 29, 177-184.

Ellson, D. G., Davis, R. C., Saltzman, J. A. & Burke, C. J. (1952). A report of research on detection of deception. (Contract N. N6 ONR 18011, Office of Naval Research). Bloomington: University of Indiana.

Fuse, L. S. (1982). Directed-lie control testing technique. Unpublished manuscript.

Gatchel, R. K., Smith, J. E., & Kaplan, N. M. (1983). The effect of propanolol on polygraphic detection of deception. Unpublished manuscript, University of Texas Health Sciences Center.

Geddes, L. A. & Newberg, D. C. (1977). Cuff pressure oscillations in the measurement of relative blood pressure. *Psychophysiology*, 14, 198-202.

Gross, J. J. & Levenson, R. W. (1993). Emotional suppression: Physiology, self-report, and expressive behavior. *Journal of Personality and Social Psychology*, 64, 970-986.

Honts, C. R. (1992). Bootstrap decision making for polygraph examinations (Grant No. N00014-92-J-1794). Grand Forks: University of North Dakota, Department of Psychology.

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.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1985). Effects of socialization on the detection of deception. *Journal of Research in Personality*, 19, 373-385.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1986). Individual differences and the physiological detection of deception. *Psychophysiology*, 23, 442. (Abstract)

Horowitz, S. W., Kircher, J. C., Honts, C. R., & Raskin, D. C. (1997). The role of control questions in the physiological detection of deception. *Psychophysiology*, 34, 108-115.

Horowitz, S. W., Kircher, J. C., & Raskin, D. C. (1986). Does stimulation test accuracy predict accuracy of polygraph tests? *Psychophysiology*, 23, 442. (Abstract)

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd Ed.). Engelwood Cliffs, NJ: Prentice Hall.

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

Kircher, J. C., & Raskin, D. C. (1981). Computerized decision-making in the detection of deception. *Psychophysiology*, 18, 204-205. (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. (1991). Manual for the Computerized Polygraph System (CPS). *Scientific Assessment Technologies*, Salt Lake City, UT, 84106

Kircher, J. C., Raskin, D. C., & Honts, C. R., (1984). Electrodermal habituation in the detection of deception, *Psychophysiology*, 21, 585 (Abstract).

Kircher, J. C., Raskin, D. C., Honts, C. R., & Horowitz, S. W. (1994). Genereralizability of statistical classifiers for the detection of deception. *Psychophysiology*, 31, S73. (Abstract)

Kircher, J.C., Woltz, D.J., Bell, B.G. & Bernhardt, P.C. (1998). Effects of audiovisual presentations of test questions during relevant-irrelevant polygraph examinations and new measures. Final report to the Central Intelligence Agency (Grant No. 110200-997-MO). Salt Lake City: University of Utah, Department of Educational Psychology.

Krapohl, D. & Norris, W. F. (2000). An exploratory study of traditional and objective scoring systems with MGQT field cases. *Polygraph*, 29, 185-194.

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

McNemar, Q. (1968). *Psychological statistics* (4th Ed.) New York: Wiley.

Numaguchi, G., Kircher, J. C., Craig, R., Raskin, D.C., Woltz, D. J., & Packard, R. E. (1994). The Erlanger method for measuring cardiovascular activity: Correlations with blood volume and arterial pressure. *Psychophysiology*, 31, S11. (Abstract)

Patrick, C. J. & Iacono, W. G. (1989). Psychopathy, threat, and polygraph test accuracy. *Journal of Applied Psychology*, 74, 347-349.

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. & Kircher, J. C. (1999) The Finapres (volume clamp) recording method in psychophysiological detection of deception examinations: Experimental comparison with the cardiograph method. *Forensic Science Communication*, 1(3), 1-17.

Podlesny, J. A. & Raskin, D. C. (1977). Physiological measures and the detection of deception. *Psychological Bulletin*, 84, 782-799.

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

Podlesny, J. A. & Truslow, C. M. (1993). Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology*, 78, 788-797.

Porges, S. W., Bohrer, R. E., Cheung, M. N., Drasgow, F., McCabe, P. M., & Keren, G. (1980). New time series statistic for detecting rhythmic co-occurrence in the frequency domain: The weighted coherence and its application to psychophysiological research. *Psychological Bulletin*, 88, 580-587.

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

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

Raskin, D. C. & Kircher, J. C. (1990). Development of a computerized polygraph system and physiological measures for detection of deception and countermeasures: A pilot study (Contract 88-L55300-000). Salt Lake City: Scientific Assessment Technologies.

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). A study of the validity of polygraph examinations in criminal investigation (Grant No. 85-IJ-CX-0040). Salt Lake City: University of Utah, Department of Psychology.

Raskin, D. C., Kircher, J. C., Horowitz, S. W., & Honts, C. R. (1989). Recent laboratory and field research on polygraph techniques. In J. C. Yuille (Ed.) *Credibility assessment*. Borerecht, The Netherlands: Kluwer Academic Publishers.

Reid, J. E. & Inbau, F. E. (1977). *Truth and deception*. Baltimore: Williams and Wilkins.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35, 651-665.

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

Spielberger, C. D., Gorsuch, R.L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA.: Consulting Psychologists Press.

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

Venables, P. H. & Christie, M. J. (1973). Mechanisms, instrumentation, recording techniques, and quantification of responses. In. W. F. Prokasy & D. C. Raskin (Eds.) *Electrodermal activity in psychological research*. New York: Academic Press.

Appendix A: Reliability Analyses

Interrater Reliability. The reliabilities of numerical evaluations performed by the original examiner and the independent evaluator are listed in Table A-1. The values in Table A-1 are Pearson product-moment correlations between the scores assigned by the original examiner and the independent evaluator. Decisions were based on 3-chart and 5-chart total scores. The reliability of scores used to make decisions exceeded .90 for the PL and DL tests.

Table A-1. Interrater reliability

	Probable-lie (n = 168)	Directed-lie (n = 168)
Skin Conductance[a]	.98	.98
Cuff Pressure[a]	.86	.83
Respiration[a]	.50	.58
Finger Pulse Amplitude[a]	.77	.64
3-chart total	.95	.91
5-chart total	.95	.93

[a]Component reliabilities were based on the three charts

Correlations Between Numerical Scores and Computer Measurements. We expected numerical evaluations to correlate with computer indices of differential reactivity because the numerical evaluators had been trained to use features that are the same or similar to those measured by the computer (Bell et al., 1999). In addition, the numerical evaluators used the Computerized Polygraph System program (CPS; Kircher & Raskin, 1991) to display the physiological recordings on the computer screen while they numerically evaluated the polygraph charts. In addition to showing the polygraph charts, CPS displays its measurements of physiological reactions on the screen as a decision aid. The availability of these computer measurements may have affected the evaluators' numerical evaluations.

The correlations between the numerical scores assigned by the independent evaluator and computer measurements are presented in Table A-2. The computer measured the amplitude of SC and cardiograph responses, and the human interpreter based his numerical scores primarily on the amplitudes of those responses. Not surprisingly, numerical scores were highly correlated with computer measurements for the SC and the cardiograph channels ($M_r > .80$). However, the computer measured respiration excursion, whereas the numerical evaluator assessed changes in respiration amplitude, cycle-time, and baseline arousal. The computer calculated the mean of its measurements of thoracic and abdominal respiration, whereas the polygraph interpreter used either the thoracic or the abdominal channel depending on which channel showed the greatest perceived difference between the comparison and relevant questions. Predictably, the correlations between computer measurements and numerical scores were lower for respiration than for SC and the cardiograph ($M_r = .57$ vs $M_r > .80$).

Table A-2. Correlations between 3-chart independent numerical evaluations and computer indices of differential reactivity

	Probable-lie (n = 168)	Directed-lie (n = 168)
Skin Conductance	.83**	.84**
Cuff Pressure	.86**	.85**
Respiration	.54**	.60**

**p < .01

Appendix B: Preliminary Analyses

Treatment-Related Attrition. Thirty-three individuals assigned to the guilty condition (16%) refused to participate after they had received their tape-recorded instructions, whereas none of the innocent subjects declined to participate. Consequently, subjects who agreed to commit the mock crime may have been sampled from a population that differed in certain respects from the more general population from which innocent subjects were drawn. Preliminary tests were conducted to explore the possibility that guilty and innocent groups differed on measures of marital status, ethnicity, occupation, age, education, or hours of sleep. We also tested if guilty and innocent subjects differed on the social-desirability scale (Crowne & Marlowe, 1964), Rotter Trust scale (Rotter, 1967), or the two anxiety scales (Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983). The guilty and innocent subjects who completed the experiment did not differ significantly on any of the demographic or personality measures.

Effects of Sex. Preliminary Guilt X Test Type X Feedback X Sex ANOVA's were conducted to test for effects of Sex on computer measures of SC amplitude, cardiograph amplitude, and respiration excursion. The Sex X Guilt interaction effect on respiration excursion was significant, F(1, 304) = 5.35, p < .02. However, since no effects of Sex were expected, only one of the 24 possible main and interaction effects involving Sex was significant, and the effect accounted for less than 2% of the variance in respiration excursion scores, Sex was dropped as a factor in all subsequent analyses.

Heterogeneity of Variance. When sample sizes are unequal, violations of the homogeneity of variance assumption can affect the risk of Type I errors (Keppel, 1991). Since the number of participants within a cell in the present study varied from 12 to 30, results of parametric statistical tests that assume homogeneity of within-cell variance were compared to the results of tests that allow for heterogeneity of variance. In all cases, the conclusions were the same. The results we report are based on the more traditional statistical tests that assume homogeneity of variance.