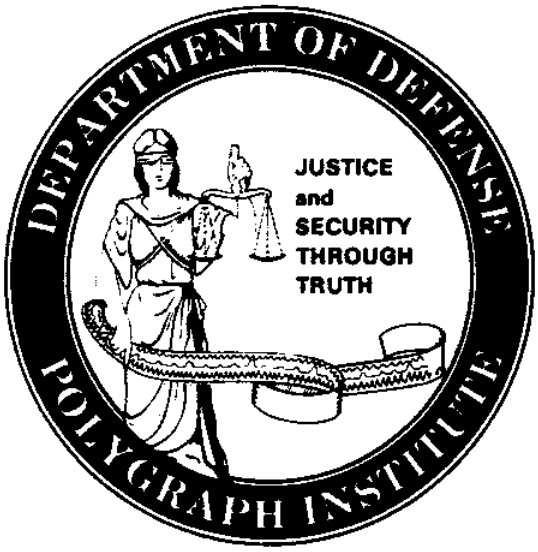
STUDIES OF THE ACCURACY OF SECURITY SCREENING POLYGRAPH EXAMINATIONS



24 March 1989

Gordon H.Barland, Charles R. Honts,
and
Steven D. Barger

RESEARCH DIVISION
DEPARTMENT OF DEFENSE
POLYGRAPH INSTITUTE

TABLE OF CONTENTS

Table of Contents

SUMMARY

The Department of Defense Polygraph Institute, founded in 1986 under Department of Defense Directive 5210.78, established a research division in January, 1987. This report describes the results of the Division's first three studies.

Using a mock espionage paradigm, Experiment 1 estimated the accuracy of the four major counterintelligence screening tests used for aperiodic security screening of persons in special access programs. Two hundred seven Army personnel and civilian employees at Ft. McClellan were given aperiodic security screening tests by examiners from the Army INSCOM, the Air Force Office of Special Investigations, the National Security Agency, and the Central Intelligence Agency. Forty-four of the 207 subjects were 'guilty' of committing acts of simulated espionage two months prior to the polygraph tests. Forty-seven subjects went through 'knowledge' scenarios in which they met someone who claimed to have committed espionage, and tried to recruit them to do likewise. Subjects in these two groups lied on the polygraph tests when they denied having committed espionage or knowing anybody who had. A third group of 116 subjects were 'innocent,' in that they were not programmed to be guilty or knowledgeable.

The polygraph examiners were very accurate at clearing the programmed innocent persons. Excluding the 3% inconclusive results, 94% of the innocent subjects were cleared. Moreover, 3 of the 6 programmed innocent subjects who were called deceptive admitted to having engaged in significant unreported real life security incidents. When those admissions were taken into account the false positive error rate was estimated to be about 3%.

There was a substantial problem in detecting the lies of the programmed-guilty and programmed-knowledgeable subjects. Excluding the 9% inconclusive results, 34% of the guilty and knowledgeable subjects were correctly identified as deceptive. The resulting false negative error rate of 66% was unexpected. Virtually no previous research has reported significant problems in detecting the lies of persons who have been programmed guilty or knowledgeable in mock crime situations. However, despite the poor identification of programmed guilty and knowledgeable subjects, approximately 20% of all subjects made admissions about real world security violations.

There are a number of possibilities as to why so many of the programmed guilty subjects were cleared. The results of Experiment 1 may reflect a reliable estimate of the actual validity of security screening tests in the field, or they may indicate the effects of some variables unique to that experimental situation. Experiments 2 and 3 tested some of the experimental variables thought to be relevant to the results of Experiment 1.

Summary

Experiment 2 compared two testing strategies for when the examiner must cover several relevant issues within the same examination. Most previous mock crime experiments have used single issue examinations, while the examinations in Experiment 1 addressed several relevant issues. Subjects in Experiment 2 were programmed to be guilty of 0, 1, 2, or 3 different acts of mock espionage or sabotage. Half of the subjects were tested with one triple issue test, and the other half were tested with three single issue tests. There was no difference in the accuracy of these two approaches to testing multiple issues. Excluding the 24% inconclusives, 79% of the innocent and 93% of the guilty subjects were correctly classified. However, neither approach was able to identify specifically which crime(s) the guilty subjects had committed. The high accuracy of the polygraph in identifying the guilty subjects in Experiment 2 was consistent with most previous research, and stands in contrast to the results of Experiment 1. Those results suggest that the false negative rate in Experiment 1 was not caused by the use of multiple issue testing or by the use mock espionage scenarios per se.

Experiment 3 examined two variables that differentiated Experiment 1 from most previous mock crime studies. First, the time interval between the commission of the mock crime and the polygraph was manipulated. In most previous mock crime experiments the polygraph examinations were administered immediately after the enactment of the mock crime. In Experiment 1 approximately two months passed between the acts of mock espionage and the polygraph examinations. In Experiment 3 half of the subjects were tested immediately following their act of mock espionage and half were tested approximately 6 weeks later.

A second variable examined in Experiment 3 concerned the specificity of the relevant questions. The relevant questions in Experiment 1 were worded in broad terms about generally defined acts of espionage and security violations. However, most studies have used very specific relevant questions. In Experiment 3 half of the subjects were asked specific relevant questions in a criminal investigative type examination and half were asked broad relevant questions in a screening type examination. The experiment was designed so that the question specificity factor was experimentally crossed with the time lag. That way the effects of the two variables could be examined both individually and in combination. Experiment 3 failed to find any effect for the manipulation of either question specificity or the experimental time lag. Excluding the 6% inconclusive outcomes, 90% of the Innocent and 81% of the Guilty subjects were correctly classified. These results suggest that neither the time lag nor the use of general relevant questions can be used as an explanation the results of Experiment 1.

One other variable discussed in the report, but not evaluated experimentally, was motivation. There was no explicit reward for subjects to produce truthful outcomes in any of these experiments. The performance of the polygraph examinations in

all three experiments was poorer than that reported in recent mock crime analog studies and a recent field study of federal examiners conducting specific issue forensic polygraph examinations. The issue of motivation in mock crime experiments needs additional research. An analysis of heart rate showed that the subjects in Experiment 1 were significantly less aroused than subjects undergoing actual screening tests in the field. It is also possible that the instructions given to the subjects in Experiment 1 could have contributed to the creation of some false negative errors. They were told that any admissions of real world security violations or criminal activity could harm their careers. That could have caused some of the programmed guilty/knowledgeable subjects to be more concerned about the control questions than about the relevant questions. If the false negative errors in Experiment 1 were caused by either low arousal or low motivation, then there should be fewer such errors in the field.

We have not yet been able to identify what factor(s) caused the high false negative rate in Experiment 1. A number of hypotheses remain to be investigated. One possibility is that the examiners' expectations of a low base rate for deception may have increased false negative errors by influencing how they conducted the pretest interview. If that is the case then Experiment 1's result's have important implications for the practice of detection of deception in the field.

The results from Experiment 1 were compared to actual data from Department of Defense polygraph examinations. Agency 4 results most closely mapped on to actual Department of Defense parameters. It appears likely that some field examiners have adjusted their testing procedures to accommodate the low base rate of deception situation they face. This adjustment is partially successful in that they make very few false positive errors. However, the analysis suggests that many deceptive individuals may be incorrectly cleared.

It may be premature to estimate the accuracy of counterintelligence screening examinations based on the evidence presently available. However, this research suggests that there are far fewer false positive errors than previously predicted and that false negative errors may be more of a problem than previously believed.

Research offers a number of possibilities for improving the accuracy of screening. High payoff areas for future research include standardization of testing techniques, data analysis, decision making, new physiological measures, and computerization of chart analysis.

ACKNOWLEDGMENTS

Only three persons are listed as the authors of this report, but more than six hundred people cooperated on the research reported here. Unfortunately, we are not able to acknowledge all of them by name. However, the Research Division's operations managers deserve special mention. They were responsible for organizing, coordinating, and supervising the scenarios and all of the polygraph examinations. Without them these projects would have been very difficult to complete. John Schwartz served as operations manager from March to June of 1987, during the planning and pilot work of Experiment 1. SA Edgar L. Stovall, Jr. has served as our operations manager since June of 1987. We wish to give special thanks to SA Stovall for his efforts. Other members of the Institute faculty also deserve thanks. They briefed the subjects on their roles, conducted hundreds of examinations, debriefed them after the tests, and helped out in countless other ways. In addition, the Institute received considerable support from Ft. McClellan and the Military Police School.

We wish to thank the examiners, supervisors, and managers from the Army, Air Force, NSA, CIA, and the Department of Defense, whose wholehearted support made this research possible. The examiners who participated in these studies had a particularly difficult task. None had participated in a scientific study before. They worked under immense psychological pressures to strive for perfection, knowing that their every move would be recorded, scrutinized, and dissected. In the real world it is difficult for a polygraph examiner to discover if he or she errs, but in the laboratory errors are disconcertingly real and immediate. Throughout these studies all of the examiners who participated showed the highest degree of professionalism.

Finally, we wish to thank the soldiers and civil servants who volunteered to undergo polygraph tests. Some were subjected to lengthy interrogations and repeated testing in the first study, yet the vast majority faithfully followed their instructions and refused to betray their assigned role to the often frustrated examiners. Months later we contacted those of the 207 subjects who had not been transferred or reassigned, to solicit their cooperation for a follow-up study. More than eighty agreed to participate again! We were continually impressed with the dedication and enthusiasm of the hundreds of colleagues and volunteers. We are profoundly indebted to them all.

GENERAL INTRODUCTION

Most research on the detection of deception has centered on the use of the control question test in criminal investigations. This research has generally indicated that the control question test is biased toward making false positive errors (see reviews by Kircher, Horowitz, & Raskin, 1988; Office of Technology Assessment [OTA] 1983; Raskin, 1988). That is, when the control question test errs, it tends to call innocent people deceptive. These reviews indicate that in high quality studies the control question test correctly classifies about 90 percent of the guilty and about 75% of the innocent subjects, although the range of validity estimates is large[1]. Considerable controversy continues in the literature regarding which studies constitute the correct data base for developing validity estimates. Additional controversy concerns the underlying rationale of the control question test. Some contend that the rationale of the control question test is completely unreasonable, and that a high false positive rate is inevitable with that technique (e.g., Lykken, 1981).

Even less is known about the accuracy of the control question test or other detection of deception techniques in screening situations. Only two studies have been reported. Correa & Adams (1981) reported 100 percent accuracy at discriminating between truthful and deceptive volunteers in a mock screening situation (N = 40). However, their validity estimate dropped to as low as 68% when the examiner had to identify which of three questions the mock guilty persons were lying about.

Barland (1981) analyzed a mock screening study designed and conducted by Steven Diduch of the 902nd MI Group. In this study, Military Intelligence examiners conducted a Counterintelligence Screening Test which utilized a directed lie control question test on 56 subjects, 30 of whom were instructed to lie to one of five items on a statement of personal history which they filled out for the experiment. The examiners cleared 76% of the programmed innocent subjects, and identified 81% of the programmed guilty subjects. As in Correa and Adams, the examiners were less accurate at identifying the precise question to which that the mock guilty subjects were lying.

Since there is so little research on the accuracy of screening tests, many scientists have reasoned from the data on criminal investigative control question tests and have argued that screening tests should make many false positive errors (Lykken, 1981; Raskin, 1984, 1986a, 1986b). These authors also

---

[1]For example, the estimates of the accuracy of control question tests with guilty criminal suspects in the OTA (1983) review ranged from 70.6% to 98.6%. The OTA estimates estimates for accuracy with innocent suspects ranged from 12.5% to 94.1%.

argue that in national security screening for spies this tendency to call innocent people deceptive will be greatly exacerbated by what is known as the base rate problem.

The base rate problem in lie detection arises when only a few examinees are lying on the test. Consider the following hypothetical example. In a population of 1000 individuals to be screened, assume that 10 are spies and 990 are innocent. Further, assume the highest estimates of validity, 95% accuracy with guilty subjects and 90% percent accuracy with innocent subjects. Given these assumptions, at least 9 of the 10 spies will be called deceptive, and 891 (90% of 990) of the innocent will be correctly cleared. However, 99 (10% of 990) of the innocent will be incorrectly called deceptive. This means that the confidence in a truthful outcome will approach 100% (891 of 892 truthful outcomes will be correct). However, the confidence in a deceptive outcome will be only 8% (only 9 of the 108 deceptive outcomes will be correct), and 92% of the deceptive outcomes will be false positive errors. This analysis, known as a conditional probability analysis, is correct as long as the underlying assumptions of base rate and test validity are correct.

Unfortunately, many of the critics of security screening end their arguments with the conditional probability analysis, implicitly suggesting that the screening process ends with the result of the polygraph examination. This type of reasoning ignores two important points. First, there are a number of safeguards built into the security screening system. Persons producing deceptive results are re-examined in order to determine why they reacted. In the event that a deceptive outcome is not resolved through testing, other types of investigation are undertaken. The ultimate impact of these safeguards on the number of false positives is not considered in a simple conditional probability analysis.

The second important point that is often ignored is the benefit gained by reducing the field of suspects. In the above hypothetical example the field of possible spies was reduced from 1,000 potential spies to a field of 108 potential spies. A screening device that can reduce the field of suspects by an order of magnitude has considerable practical utility.

There is some evidence that casts doubt on the predictions of a high false positive error rate in security screening examinations. The Department of Defense (DoD, 1986; 1987) reports to congress on the DoD polygraph program reported that of 8,581 security screening examinations conducted during 1986 and 1987, only 53 produced a deceptive outcome, and all but four of those deceptive outcomes were confirmed by admissions made by the subjects. Even if the remaining four cases were all false positive errors, the maximum possible false positive error rate would be less than 0.05%. However, the number of deceptive people who were incorrectly cleared (false negative errors) is

unknown. Thus, there is some evidence which suggests that the federal security screening polygraph program may be very good at avoiding false positive errors. These data from the Department of Defense suggest that the assumptions used by scientists in calculating their conditional probability analyses are either incomplete or in error. It may be that there are considerable validity differences in the application of criminal investigative and security screening polygraph examinations.

The reason for this apparent discrepancy between criminal investigations and screening situations is unknown. Barland (1988) has suggested that if the examiner believes that the person to be tested is almost certainly truthful, this might bias the examiner, altering the way the examination is conducted and the way the charts are interpreted. If that is true, then there would be a tendency to make it easier to clear both the innocent and guilty persons. That is, false positive errors may be reduced at the expense of increasing false negative errors.

EXPERIMENT 1

Introduction

Federal examiners use four types of examinations for aperiodic security screening of employees who already have security clearances. Since little is known about the accuracy of any of these techniques as screening tests, and since such tests are very important to national security, we decided to study those tests. Unfortunately, field research in the detection of deception is difficult, since a criterion for ground truth (who really is guilty and innocent) must be developed independently of the polygraph examination. This is difficult to do in criminal situations, but is doubly difficult in security screening where the base rate of deception is very low. Therefore, we examined security screening in an analog screening situation where we could exert experimental control over ground truth.

However, there are several problems with analog studies. The subjects of an analog study know they are participating in an experiment. This may create a number of differences between research subjects and real world examinees in terms of their reasons for taking the polygraph, the amount of stress they are under during the examinations, the type and extent of their emotional reactions, and how they react to questioning following the tests. For example, in an analog situation the programmed guilty subject answers the screening question, "Have you ever transmitted classified information to a representative of a foreign government without authorization?" with a "No." Technically, the programmed guilty subject is answering the screening question truthfully, for the confederate was not in fact representing a foreign government, and whatever information was passed was in fact authorized. This is a problem common to all mock-crime paradigms, but despite this, the detection rate for programmed guilty subjects in realistic analog mock crime studies is usually quite good.

In addition, experimental subjects volunteer, whereas the screening examinees must take the polygraph tests aperiodically to retain their clearances. This self-selection in experiments may reduce the proportion of subjects seeking to conceal real-life information from the examiner, and may reduce the generalizability of analog screening studies. However, a more serious problem for analog screening studies involves the programmed-innocent subjects and real-life information. Because the test questions encompass activities beyond the scope of the research study, it could happen that programmed innocent subjects may attempt to conceal information from the examiner regarding an actual security incident that occurred prior to the study. Thus, a deceptive outcome would be a false positive only in reference to the study scenario; it would be a true positive outcome in reality. Unless the subject volunteers information about real-life guilt, the outcome would be miscategorized as an error (false positive). Ground truth is thus harder to ascertain in

analog screening studies than in analog criminal investigative studies, especially for the programmed-innocent group.

Despite the problems associated with analog studies it was decided that they were the best way to obtain initial estimates of the validity of security screening polygraph examinations. In Experiment 1 we examined the validity of the four major federal security screening techniques with three criterion groups, Innocent, Guilty, and Knowledgeable. Innocent subjects were instructed to answer all test questions truthfully. Guilty subjects committed acts simulating espionage, and were instructed to deny having committed espionage during the polygraph examination. Knowledgeable subjects were exposed to an individual who claimed to be a spy, and were instructed to deny that knowledge during the polygraph examination.

This study introduced several methodological changes in polygraph research paradigms, in the hope of improving the generalizability of the analog paradigm. Most previous research has used only one examiner. That examiner was usually selected on the basis of availability to the experimenter rather than being randomly drawn from the target population. This study used 18 federal polygraph examiners who conduct security screening examinations on a daily basis. Although they were not randomly selected, they were believed to be generally representative of the federal screening examiner population.

Most previous analog studies have administered the polygraph tests minutes after the programmed guilty subjects have committed the mock crime. In another effort to make this analog as realistic as possible, the polygraph tests were not administered until approximately two months after the mock crimes were committed. Furthermore, the crime scenarios were more complex than most previous studies. They required several actions over a course of at least two days.

Method

Subjects

The target population for generalization of this study was those government employees with Top Secret/Codeword clearances working in special access and other programs requiring that they be aperiodically polygraphed for counterintelligence purposes. Unfortunately, such a population was not readily available at Fort McClellan, since very little Top Secret/Codeword material is handled on that post. The agencies participating in this study described the demographic characteristics of the target population, and an effort was made to select a group of civilian and military personnel at Ft. McClellan which would match the target population as closely as possible.

During April and May of 1987, volunteers were solicited during a series of small meetings (10 to 50 attendees) from among

the military and civilian personnel at Ft. McClellan, Alabama. During those meetings potential subjects were told the following: The Department of Defense wished to conduct a study on the accuracy of security screening polygraph tests. Volunteers would be asked to take a security screening polygraph examination to determine if they were security risks. Randomly selected volunteers would be asked to participate in a mock crime, which would require several hours of their time over several days. All volunteers would have to take a security screening polygraph test in August, 1987, which would take half a day. If the results were questionable, they would be asked to return for additional examinations. The tests would be conducted by federal polygraph examiners.

In an effort to simulate the anxieties of the target population, the potential volunteers were informed that any admissions of criminal activity or violations of a national security nature would be reported to an adjudication panel. If the adjudicators felt that the admissions were significant, an investigation could be opened, and their careers possibly damaged. They were told that if they had anything in their background which they didn't want to reveal, they should not volunteer for this study. Two motives for volunteering were proffered: they would be helping their country, and they would gain firsthand experience about the polygraph. No money or reward was offered either for participation or for passing the examination.

A total of 260 subjects volunteered for this study. Forty-six subjects (all military) were lost through reassignment prior to the polygraph examinations. An additional six subjects declined to participate in the guilt scenarios when they learned what was required. A total of 208 subjects were administered polygraph examinations. During the polygraph examinations, one of the subjects stated she was currently being treated for a relatively serious psychological problem that had begun after her initial briefing. Although the examiner continued the test, upon review the agency stated that its policy prohibits administering polygraph tests to persons with that type of disorder. Based on this information we removed this subject from the analysis.

Of the 207 subjects whose data were retained for analysis, 44 were programmed guilty, 47 were programmed knowledgeable, and 116 were programmed innocent. Subjects' ages ranged from 20 to 59 with a mean of 36. There were 149 males and 58 females. Sixty-six of the subjects were civilians, and 131 were military. Military ranks ranged from E4 to Colonel. The civilians were all employees of Fort McClellan.

Most subjects were asked about the highest level of classified material to which they had ever had access. Of the 181 subjects for whom data were available, 50 reported having worked with TOP SECRET material, 96 with SECRET, 8 with

CONFIDENTIAL, and 27 denied having ever had access to classified information. These data were not verified from official files.

There is one clear difference between the subjects in this study and target population for generalization. Only a quarter of the subject population is believed to have handled TOP SECRET material, whereas all of the subjects currently in the special access programs have access to TOP SECRET.

Polygraph Examiners

Eighteen polygraph examiners and five quality control supervisors were provided by the participating agencies. Each agency provided five examiners (except Agency 1 and Agency 4, which sent four) and one supervisor (except Agency 4, which provided two). Selection of the examiners was left to the agencies. However, the examiners were required to be intimately familiar with the aperiodic testing technique employed by their agency, and their primary duty had to be conducting screening examinations. Although the examiners were not randomly selected, discussions with the agencies revealed no factor which appeared to introduce any systematic bias. The selection criteria used by the agencies generally included availability (no examiner was brought in from overseas for this study) and experience (no intern examiner was selected).

The ages of the 18 examiners ranged from 27 to 47, with a mean of 38. There were 15 males and 3 females. The examiners from three of the agencies had graduated from the Defense Polygraph Institute or its forerunner, and were certified examiners. Their years of experience ranged from 1 to 13, with a mean of 5. The number of screening examinations they had conducted ranged from 225 to 2,200, with a mean of 1,092.

Apparatus

All polygraph examinations were conducted at the Defense Polygraph Institute in small, plainly furnished rooms containing two chairs and a desk. The polygraphs were recessed into the desk, with the surfaces flush. One-way observation windows allowed the quality control supervisors to monitor each examination. Two video cameras were suspended from the ceiling of each examination room and the cameras were visible to the subject. One camera viewed the subject, and the other the polygraph chart. The view of the subject was recorded throughout the examination. While the polygraph charts were being obtained, the view of the polygraph charts was recorded split-screen on the same videotape as the subject.

The examiners used standard field polygraphs that were typical of those used by their agency for security screening polygraph examinations. All of the polygraphs were manufactured either by Stoelting or Lafayette. As a minimum, each polygraph measured respiration, skin resistance, and relative blood pressure.

Procedure.

Initial Handling. Following the briefing of the potential volunteers about the purpose of the study and the hazards involved, all who agreed to participate were assigned a subject identification number and were asked to fill out brief personal history questionnaires. The information from those questionnaires allowed the scenarios to be tailored to those subjects who would be programmed guilty or knowledgeable. The subjects also filled out questionnaires concerning their attitudes toward the polygraph and how accurate they believed the polygraph to be in a variety of situations. The subjects were also administered three tests by a psychologist. None of the results of these psychological tests are reported here. This initial handling session took about three hours. The psychological testing was discontinued for the final 35 subjects because a number of potential volunteers appeared to be deterred by the testing procedure and the amount of time involved. Appendix A contains copies of the following: Consent Form, Volunteer Affidavit, Personal Data Form, Polygraph Attitudes Questionnaire, Polygraph Accuracy Questionnaire, and Subject Debriefing Questionnaire.

Scenarios. The subjects were randomly assigned to treatment conditions in a matrix of three guilt levels (innocent, guilty, and knowledgeable) and four test types. Once the number of subjects in a cell reached a predetermined ceiling, no more subjects were assigned to it. Uneven cell frequencies resulted in the final sample as the result of the attrition of subjects who were transferred from Ft. McClellan between initial assignment to conditions and the the polygraph examinations. In addition, the last several subjects originally assigned to Agency 3 were examined by Agency 4 because the testing took longer than anticipated, and the Agency 3 examiners were unable to remain at Fort McClellan.

About a week after the initial handling session, the subjects assigned to the guilty and knowledgeable conditions were individually briefed about their assignment by a member of the research staff. At that time, they were given the time, date, location, and bona fides for their initial contact in the scenario. They were not given any details of the scenario, but were directed to obey to the letter all instructions from their contact. They were told that when it came time to take their polygraph test, they would have to lie on the test. Under no circumstances were they to admit to the examiner that they had participated in a scenario. They were further told that they were not to give out any details of the scenario, even if they had to take several examinations. The subjects were not given any countermeasure instruction. Two photographs were taken of each subject so that their scenario contact would be able to identify them. The briefing, which lasted about twenty minutes, was concluded by cautioning subjects not to discuss their guilt or knowledge role with anyone, not even their spouse.

One to three days following this briefing, the guilty and knowledgeable subjects started their scenarios. A total of ten scenarios (eight guilt and two knowledge) involving mock espionage had been prepared. Each scenario was tailored to the background of the individual volunteers. For example, a black subject would not be scheduled for a contact in a bar with a predominantly white red-neck clientele.

All scenarios were planned and executed by five Air Force and INSCOM case officers who played the role of hostile intelligence agents. The scenarios all involved acts of mock espionage such as photocopying, photographing, or taking mock classified documents. None of the scenarios involved sabotage or terrorism, even though most security test formats included those areas. The guilt scenarios required two or three contacts involving one or two of the case officers over a period of two or three days. Most of the scenarios were executed during the evenings and on weekends in order to avoid arousing the suspicion of the volunteers' coworkers. This procedure was designed to avoid problems of coworkers discovering that they were both participants in the study.

Subjects assigned to the Innocent condition were not contacted again until approximately two months after the initial handling. During the second contact they were scheduled for their polygraph examination.

Experimental Controls. One of the potential problems in this type of study was the possibility that the examiners might gain unfair advantages as the experiment progressed. For example, they might have learned details of the study such as the nature of the scenarios or the base rate of deception. We attempted to retard this process with the following procedures:

(1) The examiners were not informed of the base rate for deception. Because they came from an environment where they believed the base rate for guilt to be very low, and because none had ever participated in a scientific study before, it is likely that they underestimated the number of persons programmed guilty or knowledgeable. This was supported by a survey of the examiners made late in the study.

(2) The subjects were instructed not to reveal any details of their scenarios to the examiners.

(3) The examiners were instructed not to solicit any scenario details during post-test interviews. This procedure created a problem for generalizability. Normal procedure calls for the examiner to obtain all relevant details of any admissions, followed by additional testing to determine the accuracy and completeness of a subject's explanations. We addressed this problem in the following way. When a confession was imminent, the examiner asked the subject to make a written statement while

the examiner stepped out of the room. The subject then sealed his statement in an envelope, and the examiner resumed testing on the accuracy of the statement without knowing the details. This put the examiners at a slight disadvantage as compared to field testing.

(4) The examiners were asked not to discuss their exams with other examiners.

(5) A large number of scenarios were employed, so that even if the examiners were to learn the details of one scenario and were to exchange details with other examiners, it is unlikely that they would learn all the scenarios.

Polygraph Examinations. Approximately two months after the scenarios had been enacted by the Guilty and Knowledgeable subjects, all subjects were administered a polygraph examination by an examiner from a federal agency. The examiners were instructed to follow their normal field procedures for security screening examinations. They were not to mention words such as 'experiment,' 'study,' or 'scenario.' They were instructed to work up the relevant and control questions exactly as they would in the field. They were free to run as many charts as they felt appropriate and to conduct re-examinations if they could not clear the subject on the first day. Each examiner conducted up to two exams per day. The only constraint upon the examiners was that they could not listen to the details of any confession. Confessions were handled with the procedures described above.

Blind Evaluations

Following the conclusion of all polygraph examinations, but before the results were released, the polygraph charts were submitted to quality control personnel from each of the agencies for independent blind evaluation. The independent evaluators made decisions on each relevant question and then gave a rating of each subject's overall truth and deception on an 11 point scale. Those ratings were converted to overall decisions of truthful, inconclusive and deceptive. Those overall decisions and the single question decisions were analyzed separately.

Agency Evaluations

Approximately one year following the conclusion of the polygraph examinations, evaluators from each of the agencies returned to Fort McClellan to evaluate the videotapes of the examinations to determine if the examinations conformed to each agencies' standard practice for the conduct of security screening examinations. Some of those evaluations were performed blind to the subjects' conditions, and some were performed with a knowledge of the subjects' conditions. No examiner who conducted examinations in the original data collection participated in the blind evaluations. Two questionnaires were used for the blind and non-blind evaluations and they are included as Appendix B.

Categorization of Test Outcomes The classification of test outcomes as correct or incorrect is more complex in analog screening studies than it is in investigations of mock crimes. In this study, the relevant questions were worded in general terms. A typical example would be, "Have you ever had unauthorized contact with an official or employee of a foreign government?". This generalized wording has an important implication for the classification of programmed innocent subjects, as they, or any subject, might have been deliberately concealing information pertaining to knowledge or acts prior to this study. For example, a soldier who had been stationed overseas may have improperly revealed classified information to a foreign girl friend. Since all subjects were explicitly warned that admissions would be adjudicated and could damage their careers, the soldier would not want to reveal this security violation to the examiner. Since security screening tests are designed to detect significant security compromises, a deceptive outcome verified by an admission of a serious compromise could be considered correct, despite the fact that the subject was programmed to be Innocent. We felt that if the admissions of programmed innocent subject reached a threshold of seriousness, such subjects should not be considered false positive errors.

The determination of the threshold of significance for subject admissions was necessarily arbitrary. In this study, the admissions made by programmed innocent subjects were screened for significance by a panel of three researchers. The threshold was not explicitly defined, but the following factors were considered: classification level of the compromised information, recency of the incident, and actual or potential damage to the national security. No admission made by programmed innocent subjects during the pretest interview, regardless of significance, was considered by the classification panel, since in those cases there was no clear intent to deceive the examiner. An example of a set of admissions that was considered significant was as follows: Following a deceptive outcome a subject admitted that he had discussed classified codeword information in environments where it was likely that the information had been compromised, he had taken classified material home, and that he had classified material at his home at the time of the test. There were seven programmed innocent subjects who produced deceptive outcomes. Of those seven, four made significant post-test admissions. Those four subjects are reported separately when appropriate.

Another classification requiring explanation regards programmed Guilty/Knowledgeable subjects who confessed their scenario involvement and were given a final polygraph examination to determine whether they had told the complete truth. Even though the final examination may have resulted in a decision of truthfulness, the examination was classified as having a correct outcome, since the programmed deception had been detected. Thus, for the original examiners in Experiment 1, we categorized

the accuracy of the system (the examiners' ultimate decision using all available information) rather than the more conventional blind evaluation of only the polygraph charts.

There was one case where, following an inconclusive call on the first series, the subject confessed his scenario involvement on being asked if anything was troubling him on any of the questions. This was classified as a correct outcome rather than an inconclusive, since the subject's role was correctly identified as a result of the examination procedure.

The classification rules were as follows:

S Programmed	First test Outcome	Admission	Retest Outcome	Classified
Innocent	NDI	--	--	Correct (True Negative)
Innocent	Inc or DI	No	NDI	Correct (True Negative)
Innocent	Inc or DI	No	Inc	Inconclusive
Innocent	Inc or DI	No	DI	Incorrect (False Positive)
Innocent	Inc or DI	Yes*	NDI	Correct (True Negative)
Innocent	Inc or DI	Yes*	Inc/DI	Correct but Deceptive
Gu or K	NDI	--	--	Incorrect (False Negative)
Gu or K	Inc or DI	No	NDI	Incorrect (False Negative)
Gu or K	Inc or DI	No	Inc	Inconclusive
Gu or K	Inc or DI	No	DI	Correct (True Positive)
Gu or K	Inc or DI	Yes	NDI	Correct (True Positive)
Gu or K	Inc or DI	Yes	Inc/DI	Correct (True Positive)

Gu = Guilty K = Knowledgeable Inc = Inconclusive DI = Deception Indicated

NDI = No Deception Indicated

*Programmed innocent subjects admitting to deliberately holding back relevant information about a significant real world incident.

Results[2]

At all levels of the analysis statistical tests were used to look for differences between subjects programmed Guilty and those programmed Knowledgeable. Only one of those analyses produced a significant result. Therefore, for purposes of simplicity the subsequent sections will treat the Knowledgeable and Guilty manipulations as one Guilty/Knowledgeable condition, unless otherwise noted.

Original Examiners' Classification Decisions.

All Agencies. The overall performance of the original examiners is shown in Table 1 and rate summaries are provided in Table 2. The determination of rates with Innocent subjects is somewhat difficult. Seven programmed Innocent subjects were reported as deceptive, but four of those seven subjects made significant admissions to real world security violations during their post-test interviews. Those four subjects were not used in calculating the accuracy rates reported in this and subsequent sections.

With Innocent subjects, the original examiners' classifications were 93% correct, 4% inconclusive, and 3% incorrect. Excluding inconclusives the classification of Innocent subjects was significant[3], 97% of the Innocent subjects were classified correctly, and 3% were false positive errors, $z = -9.30$, $p < 0.001$. Of the Guilty/Knowledgeable subjects, 31% were correctly identified as deceptive by the original examiners, 9% were reported as inconclusive, and 60% were incorrectly cleared. Excluding inconclusives, the classification of Guilty/Knowledgeable subjects were 34% correct, and 66% were false negative errors, $z = -2.85$, $p < 0.01$. Since the outcome falls in the opposite direction from the prediction of the appropriate alternative hypothesis for this one tailed test, this result indicates that the classification of Guilty/Knowledgeable subjects was not different from chance. However, in a practical sense, this finding means that the classification of Guilty/Knowledgeable subjects was 'significantly worse than chance.'

---

[2]Two broad types of statistics (hypothesis tests and magnitude of effect statistics) are reported in this and the subsequent results sections. Hypothesis tests (Chi Square, F, t, Binomial, and z, in this report) evaluate the outcomes of an experiment against a fixed criterion of chance. In the behavioral sciences that criterion is usually set at a probability of 0.05. That is, if the results of an experiment are likely to occur only 5 times (or less) out of 100 replications by chance, they are accepted as likely to have been caused by the independent variables of the experiment, rather than to have occurred by sampling error. It is inappropriate, and may be very misleading to interpret the probability levels of a hypothesis test as indicating the size of the effect of the independent variable. A hypothesis test result that is very unlikely by chance, $p < 0.001$, may be a smaller effect than a test with a calculated probability value that is much more likely by chance, $p < 0.05$, depending on the statistics used, and the sample size. Do not use reported probability values to evaluate the size of the effects of the variables tested. The size of the effect of a relationship between two variables is appropriately

Table 1[4]. Classification results of the 19 original examiners from all agencies.

	NDI	Inc	DI	Totals
Innocent	105	4	3 + (4)	116
Guilty	55	8	28	91
Totals	160	12	35	207

The predictive relationship illustrated in Table 1 was evaluated in several ways. First, a Chi Square ($\underline{X}^2$) analysis was conducted on the Innocent and Guilty/Knowledgeable by Decision (Truthful, Inconclusive, and Deceptive) contingency table shown as Table 1, and the resulting $\underline{X}^2$ was significant, $\underline{X}^2$ (2) = 35.33, $\underline{p} < 0.0001$. This result indicates that the decisions on Innocent and Guilty/Knowledgeable subjects were not randomly distributed across the cells of the contingency table.

Then, the tau c (Norusis, 1986) statistic was used to measure the magnitude of the predictive relationship of the original examiners' decisions for the Innocent, Guilty/Knowledgeable criteria. Tau c is a nonparametric measure of association that can range from -1.0 to +1.0 and can be interpreted in the same manner as a correlation coefficient (Siegel, 1956). We used tau c as an index of predictive performance, and as a statistic for comparison of discrimination performance between agencies.

---

evaluated with magnitude of effect statistics (r and tau C in this report). These statistics result in a coefficient value that can vary between -1.0 and +1.0. A value of +1.0 indicates a perfect direct relationship, if a value on one variable is known the value on the other variable is also known. A value of -1.0 also indicates a perfect relationship, but an inverse one. In a perfect inverse relationship, as one variable grows larger the other grows smaller at the same rate. A value of 0.0 indicates no relationship between the variables. If r is squared the resulting value represents the amount of variance the two variables share in common. Probabilities associated with r and tau C, represent the likelihood that those values could have been obtained by chance sampling.

[3]Binomial and z tests conducted on the classifications of Innocent and Guilty/Knowledgeable conditions were all conducted 1-tailed. All other statistical tests were conducted 2-tailed.

[4]The values shown in Table 1 in parentheses and boldface, (4), represent those programmed innocent subjects who made significant post-test admission that were presumed to account for their being diagnosed as deceptive.

Table 2. Summary statistics of the performance of the original examiners.

	Agency				
	1	2	3	4	ALL
N	57	52	46	52	207
INCONCLUSIVE	(2/57) 4%	(4/52) 8%	(5/46) 11%	(1/52) 2%	(12/207) 6%
RATES FOR:					
INNOCENT	(1/35) 3%	(0/28) 0%	(3/26) 12%	(0/27) 0%	(4/116) 3%
KNOWLEDGE	(1/12) 8%	(1/12) 8%	(1/10) 10%	(0/13) 0%	(3/47) 6%
GUILTY	(0/10) 0%	(3/12) 25%	(1/10) 10%	(1/12) 8%	(5/44) 11%
G + K	(1/22) 5%	(4/24) 17%	(2/20) 10%	(1/25) 4%	(8/91) 9%
INCORRECT	(11/57) 19%	(12/52) 23%	(13/46) 28%	(22/52) 42%	(58/207) 28%
RATES (EXCLUDING INCONCLUSIVES) FOR:					
INNOCENT	(0/33) 0%	(1/26) 4%	(2/22) 9%	(0/27) 0%	(3/108) 3%
GUILTY	(7/10) 70%	(4/9) 44%	(7/9) 78%	(10/11) 91%	(28/39) 72%
KNOW	(4/11) 64%	(7/11) 64%	(4/9) 44%	(12/13) 92%	(27/44) 61%
G + K	(11/21) 52%	(11/28) 55%	(11/18) 61%	(22/24) 92%	(55/83) 66%
CORRECT	(43/57) 75%	(34/52) 65%	(27/46) 59%	(29/52) 56%	(133/207) 64%
RATES (EXCLUDING INCONCLUSIVES) FOR:					
INNOCENT	(33/33) 100%	(25/26) 96%	(20/22) 91%	(27/27) 100%	(105/108) 97%
GUILTY	(3/10) 30%	(5/9) 56%	(2/9) 22%	(1/11) 9%	(11/39) 28%
KNOWLEDGE	(7/11) 64%	(4/11) 36%	(5/9) 56%	(1/13) 8%	(17/44) 39%
G + K	(10/21) 48%	(9/20) 45%	(7/18) 39%	(2/24) 8%	(28/83) 34%
CONFIDENCE IN OUTCOMES:					
DI	(10/10) 100%	(9/10) 90%	(6/9) 67%	(2/2) 100%	(27/31) 87%
NDI	(33/44) 75%	(25/36) 69%	(20/31) 65%	(27/49) 55%	(105/160) 66%
LYKKEN MEAN ACCUARCY	((100%+48%)/2) 74%	((96%+45%)/2) 70%	((91%+39%)/2) 65%	((100%+8%)/2) 54%	((97%+34%)/2) 66%

The tau c for all agencies combined was significant, tau c = 0.34, p < 0.0001, but was of a modest magnitude. In comparison, the tau c derived from a recent mock crime study (Kircher & Raskin, 1988) was 0.87. Similarly, the original examiners in a field study of specific issue forensic polygraph examinations conducted by the United States Secret Service produced a tau c of 0.76 (Honts, Raskin, Kircher, & Horowitz, 1988).

Finally, effects in classification performance of the combined agencies were tested with a series of Kruskal-Wallis nonparametric analyses of variance (ANOVA). A significant difference in classification of Innocent and Guilty/Knowledgeable subjects was indicated by the first Kruskal-Wallis ANOVA, χ^2 (1) = 17.50, p < 0.0001. This effect indicates that Innocent and Guilty/Knowledgeable subjects were classified differently. A second Kruskal-Wallis ANOVA indicated that there were significant differences in classification performance between the agencies, χ^2 = 5.61, p < 0.05.

An examination of the performance of the individual agencies suggested that there might be an interaction of Agency and Condition in the decision data. Since there is no nonparametric statistical test for interaction effects, a parametric Condition (Guilty/Knowledgeable, Innocent) by Agency (CIA, MI, NSA, OSI) ANOVA was conducted to investigate the possibility of interaction. This may violate the assumption of interval scale measurement of a parametric ANOVA. However, Kircher, Horowitz, and Raskin (1988) have argued that the decisions NDI, Inconclusive, DI represent an interval scale, and have used parametric statistics on such data. For purposes of this analysis and when necessary in some additional analyses, we also treated those decisions as an interval scale. ANOVA produced results that were similar to the Kruskal-Wallis ANOVAs, with significant main effects for Condition, F (1, 195) = 47.65, p < 0.001, and Agency, F (3, 195) = 4.85, p < 0.01. ANOVA also indicated a significant interaction of Condition and Agency, F (3,195), p < 0.05. That interaction appears to be primarily due to the very poor performance of Agency 4 with Guilty/Knowledgeable subjects.

Agency 1. The classification results of the original examiners of Agency 1 are shown in Table 3, and they are summarized in Table 2. With Innocent subjects, the Agency 1 examiners' classifications were 97% correct, 3% inconclusive, and 0% incorrect, z = -5.31, p < 0.001. With Guilty/Knowledgeable subjects, the Agency 1 examiner's decisions were 46% correct, 4% inconclusive, and 50% incorrect. Excluding inconclusives the classifications of Guilty/Knowledgeable subjects were not different from chance, 48% of those decisions were correct, and 52% were false negative errors, Binomial, p < 0.50, ns. The χ^2 for the Agency 1 decision table was significant, $\chi^2(2)$ = 19.32, p < 0.001, as was the tau c, tau c = 0.46, p < 0.001.

Table 3. Classification results of the original examiners from Agency 1.

	NDI	Inc	DI	Totals
Innocent	33	1	(1)	35
Guilty	11	1	10	22
Totals	44	2	11	57

Agency 2. The classification results of the original examiners of the Agency 2 are shown in Table 4, and they are summarized in Table 2. With Innocent subjects, the Agency 2 examiners' classifications were 96% correct, 0% inconclusive, and 4% incorrect, $z = -4.53$, $p < 0.001$. With Guilty/Knowledgeable subjects, the Agency 2 examiners' decisions were 37% correct, 17%

Table 4. Classification results of the original examiners from Agency 2.

	NDI	Inc	DI	Totals
Innocent	25	0	1 + (2)	28
Guilty	11	4	9	24
Totals	36	4	12	52

inconclusive, and 46% incorrect. Excluding inconclusives the classifications of Guilty/Knowledgeable subjects were not different from chance, 45% of those decisions were correct, and 55% were false negative errors, Binomial, $p = 0.25$, ns. The X^2 for the Agency 2 decision table was significant, X^2 (2) = 15.79 $p < 0.001$, as was the tau c, tau c = 0.50, $p < 0.001$.

Agency 3. The classification results of the original examiners of Agency 3 are shown in Table 5, and the rate data are summarized in Table 2. With Innocent subjects, the Agency 3 examiners' classifications were 77% correct, 11% inconclusive, and 12% incorrect. Excluding inconclusives the classification rates for Innocent subjects were significant, 87% of the decisions with Innocent subjects were correct and 13% were false positive errors, Binomial $p < 0.001$. With Guilty/Knowledgeable subjects, the Agency 3 examiners' decisions were 35% correct, 10% inconclusive, and 55% incorrect. Excluding inconclusives, the classification of Guilty/Knowledgeable subjects was not different from chance; 39% of those decisions were correct, and 61% were false negative errors, Binomial, $p = 0.24$, ns. The X^2 for the Agency 3 decision table was not significant, but the the tau c was significant although modest, tau c = 0.28, $p < 0.05$.

Table 5. Classification results of the original examiners from Agency 3.

	NDI	Inc	DI	Totals
Innocent	20	3	2 + (1)	26
Guilty	11	2	7	20
Totals	31	5	10	46

Agency 4. The classification results of the original examiners of Agency 4 are illustrated in Table 6 and they are summarized in Table 2. With Innocent subjects, the Agency 4 examiners' classifications were 100% correct, 0% inconclusive, and 0% incorrect, $z = -5.00$, $p < 001$. With Guilty/Knowledgeable subjects, the Agency 4 examiners' decisions were 8% correct, 4% inconclusive, and 88% incorrect. Excluding inconclusives, 8% of the decisions with Guilty/Knowledgeable subjects were correct, and 92% were false negative errors, Binomial, $p < 0.001$. This result could be interpreted to indicate that the classification of Guilty/Knowledgeable subjects was 'significantly worse than chance.' X^2 was not calculated for the Agency 4 decision table since the expected frequencies were so low. The tau c, although very modest, was significant, tau c = 0.12, $p < 0.05$.

Table 6. Classification results of the original examiners from Agency 4.

	NDI	Inc	DI	Totals
Innocent	27	0	0	27
Guilty	22	1	2	25
Totals	49	1	2	52

Original Examiners' Confidence in Outcomes.

Analysis of Variance (ANOVA) was used to examine differences in the ratings examiners gave on the confidence scales. Those ratings were subjected to a subject Gender (male/female) X Condition (Innocent, Knowledgeable, Guilty) X Agency ANOVA. None of the second or third order interactions were significant. The main effect for Gender was significant, F (1, 184) = 4.58, $p < 0.05$, indicating that the examiners were more confident in their decisions on females (M = 4.27) than on males (M = 3.85). However, a Kruskal-Wallis ANOVA and a parametric ANOVA failed to find any difference in decision rates between genders, nor did gender interact with guilt condition in the decisions.

Examiners were more confident in their decisions on Innocent (M = 4.12) and Knowledgeable subjects (M = 4.04) than they were in their decisions on Guilty subjects (M = 3.50) as was indicated by a significant main effect for Condition, F (2, 184), = 6.03, $p < 0.05$. This finding suggests that examiner confidence in an outcome has little relationship to the accuracy of the outcome, since examiners were very accurate with Innocent subjects, but not very accurate with Knowledgeable or Guilty subjects. That hypothesis was explored by coding outcomes as Correct, Inconclusive, and Incorrect, and then correlating those codes with confidence in outcome. The resulting correlation was significant, r = -0.12, , $p < 0.05$, but indicated that confidence and outcome share only 1.4% common variance. Agency 2 examiners were more confident in their decisions (M = 4.46) than were Agency 3 (M = 3.70), Agency 1 (M = 3.95), or Agency 4 (M = 3.75) examiners as was indicated by a significant main effect for Agency, F (3, 184) = 4.20, $p < 0.01$.

Real World Admissions by Subjects.

The number and seriousness of real world admissions given by subjects during their examinations is illustrated by agency in Table 7. A Kruskal-Wallis ANOVA indicated that there was a significant difference between the agencies in the number of real world admissions they obtained, χ^2 (1) = 17.89, $p < 0.001$. Agency 3 obtained the most real world admissions and Agency 4 the least. However, this is not at all surprising. Most of the real world admissions obtained were of security violations. Agency 4 policy is that they are chartered to detect espionage and sabotage, and that they are not chartered to search for and or report security violations. Since Agency 4 examiners neither look for nor report security violations we should not expect them to obtain many admissions to security violations.

Table 7. Percent real world admissions by severity and agency.

	Admission Severity				
Agency	None	Petty	Minor	Moderate	Significant
Agency 1	81	12	5	2	0
Agency 2	90	4	0	4	2
Agency 3	68	15	11	4	2
Agency 4	96	2	2	0	0

Real world admissions were also examined in terms of whether they were obtained from subjects who were programmed Innocent, Knowledgeable, or Guilty. A Kruskal-Wallis ANOVA found no significant difference between the conditions.

Blind Evaluations

The confidence ratings of overall truthfulness made by the independent evaluators two months after the examinations are shown in Table 8. Those ratings were subjected to a series of a priori contrasts to see if the ratings for Innocent subjects were different from those of the Knowledgeable and the Guilty groups. The a priori contrasts indicated that the ratings for the Innocent group were different from those of the Knowledgeable group, t (198) = 2.7, $p < 0.01$, and from the combined Knowledgeable and Guilty groups, t (198) = 2.83, $p < 0.01$, but the ratings of the Innocent group were not different from the Guilty group.

Table 8. Mean Ratings of Truth and Deception by the Independent Evaluators.

Condition	Mean Rating (-5 = Deceptive, +5 = Truthful)
Innocent	1.43
Knowledgeable	0.04
Guilty	0.49
Guilty/Knowledgeable	0.26

A decision analysis of the overall truthfulness ratings was conducted by converting the confidence ratings to decisions with varying inconclusive zones. Initially, ratings greater than 0 were considered truthful, ratings less than 0 were considered deceptive, and 0 ratings were considered inconclusive. The inconclusive zone was varied by one rating point in each direction until the inconclusive zone was -4 to +4 inclusive. The predictive validity of the decisions made this way was significant and peaked at a tau c value of 0.24, $p < 0.05$, when the inconclusive zone was 0. A decision table was created using this 0 inconclusive zone for all examinations and all agencies and is presented as Table 9. Of the individual agencies only Agency 1 produced a significant discrimination of Innocent and Guilty/Knowledgeable subjects.

Table 9. Percent Decisions Based on Blind Evaluators' Ratings

Agency	Correct Innocent	Inc. Innocent	Incorrect Innocent	Correct Guilty	Inc. Guilty	Incorrect Guilty	tau c
Agency 1	71	12	17	64	9	27	0.47*
Agency 2	36	28	36	64	9	27	0.23
Agency 3	70	4	26	43	14	43	0.25
Agency 4	93	7	0	4	16	80	0.13
Combined	67	13	20	42	12	46	0.24*

* $p < 0.05$

Accuracy on Single Relevant Questions.

The original examiners' and the blind evaluators' decisions on single relevant questions were also evaluated. We have decisions from the original examiners on 1194 relevant questions. The accuracy of the original examiners on single relevant questions is shown in Table 10. With truthfully answered relevant questions, the independent evaluators were correct 88% of the time, incorrect 2% of the time, and called 10% inconclusive. Excluding inconclusives, 97% of the calls on truthfully answered questions were correct and 3% were false

Table 10. Classification of single relevant question by the original examiners from all of the agencies.

	NDI	Inc	DI	Totals
Innocent	766	82	21	869
Guilty	206	54	65	325
Totals	972	136	86	1194

positive errors, z = -27.0, $p < 0.001$. When questions were answered deceptively, the independent evaluators were correct 20% of the time, incorrect 63% of the time, and called 17% inconclusive. Excluding inconclusives, 24% of the deceptively answered questions were correctly classified and 76% were false negative errors, z = -8.51, , $p < 0.001$. Again, this could be interpreted as performance that was 'significantly below chance.' The discrimination between truthfully and deceptively answered relevant questions was significant, tau c = 0.21, $p < 0.001$, but was modest in magnitude. A similar analysis was performed for each agency and is summarized in Table 11.

Table 11. Percent Decisions On Single Relevant Questions by the Original Examiners

Agency	Correct Truthful	Inc. Truthful	Incorrect Truthful	Correct Deceptive	Inc. Deceptive	Incorrect Deceptive	tau c
Agency 1	86	11	3	18	24	58	0.22**
Agency 2	90	7	3	24	13	63	0.22**
Agency 3	79	19	2	27	18	55	0.20**
Agency 4	95	4	1	4	12	84	0.09*
Combined	88	9	3	20	17	63	0.21**

* $p < 0.05$ ** $p < 0.01$

The independent evaluators' decisions were also evaluated at the level of accuracy on single relevant questions. We have evaluations for the blind evaluators for 1318 relevant questions. The accuracy of the blind evaluators on single relevant questions is shown in Table 12. With truthfully answered relevant

Table 12. Classification of single relevant question by the independent evaluators from all of the agencies.

	NDI	Inc	DI	Totals
Truthful	652	231	72	955
Deceptive	164	150	49	363
Totals	816	381	121	1318

questions, the independent evaluators were correct 68% of the time, incorrect 8% of the time, and called 24% inconclusive. Excluding inconclusives, 91% of the calls on truthfully answered questions were correct and 9% were false positive errors, $z = -21.52$, $p < 0.001$. When questions were answered deceptively, the independent evaluators were correct 14% of the time, incorrect 45% of the time, and called 41% inconclusive. Excluding

inconclusives, 23% of the deceptively answered questions were correctly classified and 77% were false negative errors, z = -7.81, $p < 0.001$. The discrimination between Innocent and Guilty subjects was significant, tau c = 0.19, $p < 0.05$. A similar analysis was performed for each of the agencies and is summarized in Table 13. With single questions only Agency 3 failed to discriminate at a significant level. Agency 1 produced the most accurate decisions on single relevant questions.

Table 13. Percent Decisions On Single Relevant Questions by the Independent Evaluators

Agency	Correct Innocent	Inc. Innocent	Incorrect Innocent	Correct Guilty	Inc. Guilty	Incorrect Guilty	tau c
Agency 1	74	24	2	10	56	34	0.33*
Agency 2	62	28	10	18	40	42	0.15*
Agency 3	52	33	15	16	40	44	0.06
Agency 4	92	7	1	6	30	64	0.21*
Combined	68	24	8	14	41	45	0.19*

* $p < 0.05$

Agency Evaluations of the Examination Materials

Blind Evaluations. Evaluators from the agencies reviewed the case materials 1 year after the conclusion of the experiment to determine if the examinations were conducted according to their agency's standards. The majority of their responses to the items in the Blind Evaluation Questionnaire (Appendix B) fell around the 'about the same' response. An interesting finding was that the blind evaluators seemed to feel that the examinations they viewed would be more accurate if the subjects were Innocent (M = 5.4), than if they were Guilty/Knowledgeable (M = 4.6), t (43) = 3.30, $p < 0.01$. ANOVA was used to test for differences between agencies in their responses to items. The mean item ratings where ANOVA indicated significant differences between agencies are summarized in Table 14, and the mean item ratings where there were no differences between agencies are summarized in Table 15 (Due to a lack of available personnel, Agency 2 did not participate in the blind evaluations).

Non-Blind Evaluations. Generally, the ratings for the questionnaires were similar to the blind evaluations and fell on or around the 'about the same' or 'just like the field' choices of the scales. That is, the behavior of the original examiners was generally evaluated as not different from standard field

practice. The ratings by the non-blind evaluators are summarized for significant differences between agencies in Table 16 and for no significant differences between agencies in Table 17.

Table 14. Mean Ratings of the Blind Evaluators With Significant Differences Between Agencies.

	Agency 1 (N)	Agency 3 (N)	Agency 4 (N)	F (df)
PRETEST LIKE THE FIELD? (7=JUST LIKE IT)	3.9 (14)	5.5 (6)	4.9 (24)	7.40, $p < 0.01$ (2, 41)
EXAMINER'S DESCRIPTION OF POLYGRAPH (4=ABOUT THE SAME)	3.6 (14)	3.8 (6)	4.8 (24)	7.90, $p < 0.01$ (2, 41)
ADMONITIONS ABOUT MOVEMENT (4=SAME AS FIELD)	4.5 (14)	4.3 (6)	3.7 (24)	7.90, $p < 0.01$ (2, 41)
SIMILAR PRESENTATION OF RELEVANT? (7=JUST LIKE THE FIELD)	3.9 (14)	4.7 (6)	5.5 (24)	13.40, $p < 0.001$ (2, 41)
SIMILAR PRESENTATION OF CONTROL? (7=JUST LIKE FIELD)	4.0 (14)	5.3 (6)	5.3 (24)	6.90, $p < 0.01$ (2, 41)

ALL $p < 0.05$

Table 15. Mean Ratings of the Blind Evaluators
No Significant Differences Between Agencies.

	Agency 1	Agency 3	Agency 4
LENGTH OF PRETEST	4.3	4.2	4.3
	(14)	(6)	(24)
(4=ABOUT THE SAME AS FIELD)			
EMPHASIS ON RELEVANT TYPICAL?	4.0	4.5	4.1
	(14)	(6)	(24)
(4=ABOUT THE SAME AS FIELD)			
EMPHASIS ON CONTROL TYPICAL?	3.9	3.8	4.0
	(14)	(6)	(24)
(4=ABOUT THE SAME AS FIELD)			
IF GUILTY, PRODUCE ACCURATE OUTCOME?	4.8	4.8	4.4
	(14)	(6)	(24)
(7=VERY ACCURATE)			
IF INNOCENT, PRODUCE ACCURATE OUTCOME?	5.4	4.8	5.5
	(14)	(6)	(24)
(7=VERY ACCURATE)			
EMPHASIS ON RELEVANTS BETWEEN CHARTS	4.7	3.3	4.0
	(3)	(3)	(4)
(4=ABOUT THE SAME)			
EMPHASIS ON CONTROLS BETWEEN CHARTS	4.0	3.7	4.0
	(3)	(3)	(4)
(4=ABOUT THE SAME)			

ALL $p > 0.05$, ns

(N)

Table 16. Mean Ratings of the Non-Blind Evaluators With Significant Differences Between Agencies.

	AGENCY				
	1 (N)	2 (N)	3 (N)	4 (N)	F
PRETEST LIKE THE FIELD? (7=JUST LIKE THE FIELD)	5.7 (9)	6.5 (15)	4.8 (15)	5.4 (12)	4.30, $p < 0.01$ (3, 47)
DESCRIPTION OF POLYGRAPH TYPICAL? (7=MORE THAN THE FIELD)	4.4 (9)	6.0 (15)	3.8 (15)	4.0 (12)	7.20, $p < 0.001$ (3, 47)
PRESENTATION OF RELEVANT THE SAME? (7=JUST LIKE THE FIELD)	5.9 (9)	6.6 (15)	5.0 (15)	5.5 (12)	4.70, $p < 0.01$ (3, 47)
HOW WELL RELEVANT QUESTIONS COVERED SCENARIO (7=COVER COMPLETELY)	6.0 (9)	6.8 (15)	4.7 (15)	3.8 (12)	13.10, $p < 0.001$ (3, 47)
DID CONTROLS OVERLAP SCENARIO? (7=NO OVERLAP)	6.9 (9)	--- (0)	4.8 (15)	3.1 (12)	29.10, $p < 0.001$ (2, 33)

Table 17. Mean Ratings of the Non-Blind Evaluators
With No Significant Differences Between Agencies.

	Agency 1	Agency 2	Agency 3	Agency 4
LENGTH OF PRETEST	4.2	3.8	3.9	4.3
	(9)	(15)	(15)	(12)
(4=ABOUT THE SAME)				
PRESENTATION OF CONTROL SAME?	5.9	0.0	4.9	5.3
	(9)	(0)	(15)	(12)
(7=JUST LIKE THE FIELD)				
EMPHASIS ON RELEVANT IN PRETEST	3.9	0.0	4.0	4.0
	(9)	(0)	(15)	(12)
(4=ABOUT THE SAME)				
EMPHASIS ON CONTROLS IN PRETEST SAME?	4.0	0.0	4.0	3.8
	(9)	(0)	(15)	(12)
(4=ABOUT THE SAME)				
ADMONITIONS ABOUT MOVEMENT?	4.0	0.0	3.9	3.8
	(9)	(0)	(15)	(12)
(4=SAME AS FIELD)				
EMPHASIS ON RELEVANTS BETWEEN CHARTS	3.7	0.0	3.7	0.0
	(9)	(0)	(11)	(0)
(4=ABOUT THE SAME AS FIELD)				
EMPHASIS ON CONTROLS BETWEEN CHARTS	4.1	0.0	4.0	0.0
	(9)	(0)	(11)	(0)
(4=ABOUT THE SAME AS FIELD)				

ALL $p > 0.05$, ns

(N)

The Agency 2 evaluator rated the examiners' descriptions of the polygraph and related physiology as being more detailed than the field, and the Agency 2 rating on this item was significantly higher than the other agencies, F (3, 47) = 7.2, $p < 0.001$. Agency 4 rated the relevant questions significantly worse at covering the scenario than the other three agencies (see Table 16). The differences between agencies on their ratings of emphasis of the relevants on the pretest were statistically significant, but do not differ greatly in magnitude. Agency 1 rated the control questions' overlap as almost none compared to the other agencies, who rated the overlap as somewhere between all or none.

Comments By the Agency Evaluators. The dominant theme expressed by the Agency 4 evaluators was a concern that the scenario situation did not have enough significance for the subject. A salient example offered was that the subject was a loyal officer and that his participation in the study was a constructive effort for his country. On the other hand, the control questions may have been perceived to be a greater threat, as questions about this person's overall honesty and integrity were more threatening than those relating to the scenario. In particular, the 'security control' questions (i.e., Have you ever discussed classified information over the telephone?) were thought to have been more relevant (as a result of real world experiences) than the programmed espionage[5]. Various cues from the examiners ('..this is just a scenario,' the exam is for 'security suitability,' etc.) were seen by the evaluators as an opportunity for rationalization on the part of the subject.

The issue of availability for follow-up testing also seemed to be important to the Agency 4 evaluators. In one case it was the opinion of the evaluator that the subject was not fit to be tested at the time of the examination, although the original examiner tested him anyway. Other distractions also contributed to taking the focus off of the exam (i.e., the subject had other appointments after the polygraph examination).

The Agency 3 reviews revealed flaws in certain tests. Against standard practice, one examiner gave instructions on how to control breathing, another emphasized the irrelevant questions, and a third interrogated on the control questions. However, the agency evaluators did not feel that these violations of standard practice invalidated any of the examinations.

---

[5]As a result of these comments, the hypothesis that control questions of a 'security nature' were too strong for a mock espionage experiment was tested in the Agency 4 data. Approximately half of the Agency 4 Guilty/Knowledgeable subjects were asked one or two 'security' control questions while the other half were asked no 'security' control questions. Statistical analysis failed to reveal any association of the 'security' control questions with decisions at either the end of the first series of questions or at the conclusion of all testing.

The Agency 2 reviews tend to focus on inadequate procedures on the part of the examiners (brief pretests, incomplete elaboration of relevant questions, etc.). There were, however, equally positive comments throughout. It appears that motivation on the part of the subjects was the major concern of the evaluators.

Questionnaire Data

The results of the analysis of the questionnaire data is presented in detail in Appendix C. There were two interesting findings. First, subjects in Experiment 1 described their emotional state during the examinations as being curious and hopeful rather than as being fearful, tense, or nervous. Second, pretest perceptions of how accurate polygraph tests were did not have any predictive validity for the outcome of the examination. That is, subjects who before the examination thought the polygraph did not work were just as likely to be correctly classified as those who thought the polygraph was very accurate. This finding does not support those critics who state that a belief in the accuracy of the polygraph is necessary for the technique to work (Lykken, 1981). Please see Appendix C for details of these analyses.

DISCUSSION

The results of Experiment 1 were surprising. In contrast to most of the scientific literature on the detection of deception, very few false positive errors and many more false negative errors were found. The bias for passing individuals was so strong that the overall performance on Guilty/Knowledgeable subjects was 'significantly poorer than chance,' and no individual agency performed at better than chance levels with Guilty/Knowledgeable subjects. The predictive validity of the screening examinations in this study was so poor that performance was at or near chance levels for two of the agencies. By way of comparison, the original examiners in a recent study of the forensic polygraph examinations given by the United States Secret Service (Honts et al. 1988; also reported as Raskin, Kircher, Honts, & Horowitz, 1988) accounted for more than six times the amount of variance, and the blind evaluator in a recent mock crime study (Kircher & Raskin, 1988) accounted for 8 times the variance in the Guilt/Innocence criterion than did the original examiners in the present study. The independent evaluators in Experiment 1 generally performed about the same as the original examiners, with the exception of the independent evaluators of Agency 2 who performed at less than half the efficiency of their original examiners (tau cs of 0.23 and 0.50, respectively).

The major unresolved question about Experiment 1 is whether the high false negative rate generalizes to security screening in the field, or whether it was an artifact of the experimental conditions. If the results do represent the field accuracy of security screening examinations, then there must be major differences, as yet undefined, between security screening and forensic polygraph examinations. In that case, it is necessary to determine what those differences are, and how their effects can be counteracted.

However, it may be that the false negative rate obtained in Experiment 1 does not generalize to the field. In order to explore that possibility, we examined the execution of Experiment 1, and the methodological differences between Experiment 1 and studies conducted in other laboratories. In that way any important methodological problems with Experiment 1 should become evident.

One issue that might be raised concerns the extent to which the techniques used in Experiment 1 actually reflect those techniques used in security screening polygraphs given in the field. That issue has been examined and can be dismissed as a possible flaw in Experiment 1. The evaluators from the agencies did not find major differences between the way examinations were conducted in Experiment 1 and the way they were conducted in the field by the respective agencies. However, there were several differences between the methodology used in this study and that used in many of the other simulation studies of the detection of deception.

The first difference concerns the number of issues in the examination. The screening examination is a multiple issue examination, while most forensic examinations are single issue examinations. Few research studies have examined multiple issue testing and those that have, have produced results that suggest a decrease in predictive validity when multiple issue examinations are used, particularly if the subject is truthful to some issues but deceptive to others (Podlesny & McGhee, 1987; Raskin, Kircher, Honts, & Horowitz, 1988). Experiment 2 was designed to examine multiple issue testing. Subjects guilty of none, 1, 2, or 3 mock crimes were tested with either one multiple issue test, or with three single issue tests. The results of Experiment 2 should provide some insight into the effects of testing multiple relevant issues within the same examination.

A second difference between Experiment 1 and most other analog studies of the detection of deception concerns the delay between the enactment of the mock crime and the polygraph examination. Experiment 1 imposed a delay of about 2 months between the enactment of the espionage scenario and polygraph examination. Most other simulation studies have imposed no delay between the enactment of their mock crime and their examinations. A few recent studies (Honts, Hodes, & Raskin, 1985; Honts, Raskin, & Kircher, 1986; 1987; Podlesny & McGhee, 1987) had a one week delay between the mock crime and the polygraph testing. All of those studies have produced results comparable to other high quality studies in the literature that did not include a time delay. We examined the effects of a lengthy time delay on security screening examinations in Experiment 3. In Experiment 3 some subjects were tested immediately after committing an act of mock espionage, and other subjects were tested 6 weeks later.

A third difference between Experiment 1 and most other simulation studies of the detection of deception concerns the specificity of the relevant questions. The relevant questions used in the security screening examinations of Experiment 1 were worded in very general terms about committing unspecified security violations. Typically, in forensic polygraph examinations very specific relevant questions are used that deal with a single well defined act. It may be that the use of nonspecific relevant questions makes it easier for deceptive individuals to produce truthful outcomes. We examined the effects of using specific and non-specific relevant questions in Experiment 3. Some subjects received typical screening non-specific relevant questions. Other subjects received very specific questions about the scenario, questions much more like those typically used in forensic polygraph examinations. Experiment 3 was designed so that the specificity of the relevant questions was crossed with the time delay and guilt/innocence in a 2 X 2 X 2 factorial design. By crossing the three factors their possible interactions could also be examined.

An important difference between Experiment 1, many other simulation studies of the detection of deception, and actual security screening examinations concerns motivation. In Experiment 1, the subjects were told that any real world admissions they made could be used against them, and guilty subjects were told that they should attempt to appear truthful and should not confess. However, there were neither benefits to the subjects if they passed their tests, nor penalties if they failed them. There is some evidence that the subjects in Experiment 1 were less aroused physiologically then were subjects in actual screening examinations. Heart rate data were calculated from the beginning of each subjects' charts in Experiment 1, and a mean heart rate was calculated for all subjects, $\underline{M}$ = 75.9. Heart rate data was also obtained from 412 individuals who took actual aperiodic screening examinations at Agency 2, $\underline{M}$ = 83.7. The difference between the heart rates in Experiment 1 and the Agency 2 subjects was significant, $\underline{F}$ (1, 616) = 34.88, , $\underline{p} < 0.001$.

Some research has suggested that motivation is an important variable in conducting simulation studies of the detection of deception. A recent meta analysis (Kircher, Horowitz, & Raskin, 1988) found that about 53% of the variance between the accuracy rates of simulation detection of deception studies was accounted for by level of motivation, with higher motivation producing more accurate results. Any differences in performance between Experiment 1 and Experiments 2 and 3 might provide some insight on this issue since all three experiments used the same level of motivation. We will return to the issue of motivation in the General Discussion section of this report.

Experiment 2

INTRODUCTION

Screening tests and criminal investigative tests differ in the number of issues they cover. Criminal investigative tests are usually limited to one specific issue ("Did you steal that money?") or to a cluster of closely related issues ("Do you know who stole that money?" "Did you steal that money?" "Do you know where any of that stolen money is now?"). On the other hand, screening tests may cover several security issues, such as espionage, sabotage, or terrorism, and a variety of lifestyle issues.

Few studies have examined the accuracy of the polygraph in multiple issue testing situations and only two studies (Barland, 1981; Correa & Adams, 1981) dealt explicitly with screening situations. In general, these studies found that polygraph examinations were more accurate at discriminating completely truthful subjects from subjects who were attempting deception to something, than at the more difficult task of discriminating to which question(s) a person was attempting deception.

Of the studies that have been concerned with multiple issues, only the study reported by Barland (1981) used federal polygraph screening procedures. That study examined the validity of Counterintelligence Screening Tests (a directed lie control test) in an analog experiment that used 56 INSCOM volunteers as subjects. Those subjects filled out a statement of personal history. Later, the 30 subjects assigned to the guilty group filled out a second statement of personal history, and they were required to lie to one of five items on this second statement. They were also instructed to lie to the same item on their subsequent polygraph test. All subjects were then tested by INSCOM examiners. Excluding the 16% inconclusive outcomes, 76% of the programmed innocent subjects and 81% of the programmed guilty subjects were correctly classified by those examinations. Decisions about deception to single questions were less accurate. Excluding the 15% inconclusive outcomes, 91% of the decisions on the questions answered truthfully, but only 63% of the questions answered deceptively, were correctly classified.

Unfortunately, there are several factors in the Barland study that may limit its generalizability. First, the guilty subjects never attempted deception to more than one relevant question. In the field it is likely that persons engaged in espionage would have to attempt deception to several relevant questions. Second, the deception was related to falsification of a statement of personal history, rather than toward the usual issues of aperiodic screening examinations. Finally, the testing technique used in the Barland study is used only by INSCOM.

Experiment 2 examined accuracy when the guilty subjects may have been lying to any one, any two, or all three issues on a

three issue test. Mock espionage and sabotage paradigms were used. The relevant issues of the examinations were similar to those used in screening examinations. The testing technique was a control question test, the technique most commonly used in criminal investigations. In addition, Experiment 2 compared the accuracy of a single multiple issue test to the accuracy of three single-issue tests.

Method

Subjects

The Subjects were 100 basic trainees at Ft. McClellan, Alabama who volunteered for the study. No pay or inducements were given to the trainees for volunteering, nor were they offered any reward for passing their polygraph examinations. They ranged in age from 18 to 32 with a mean of 20.2 years. Ninety-four of the subjects were males and 6 were females.

Apparatus

Lafayette all-electronic field polygraph instruments were used. Those instruments recorded respiration by means of an elastic, air-filled tube placed around the subject's chest. Relative blood pressure was measured by means of an arm cuff inflated to about 70 mm Hg placed on the subject's upper right arm. Vasomotor activity was measured by means of a photoelectric plethysmograph placed on the subject's left thumb. Skin resistance was measured by stainless steel plate electrodes attached to the palmar surface of the subject's left index and ring fingers. Skin conductance was measured by stainless steel plate electrodes attached to the palmar surface of the subject's left middle and little fingers. No electrolyte medium was used for either skin resistance or conductance measurement. The examinations were administered in the same exam rooms described in Experiment 1. All of the examinations were videotaped using procedures similar to those described for Experiment 1.

Procedure

Subjects were randomly assigned to one of four conditions of equal size. One condition was an innocent condition and the other three were guilty conditions. Subjects assigned to the first guilty condition enacted one of three possible acts of espionage or sabotage. Subject assigned to the second guilty condition enacted two of the three possible acts, and the remaining guilty subjects enacted all three mock crimes.

Subjects were brought to the Polygraph Institute from their training area in groups of six to ten. They were briefed, as a group, on the purpose of the experiment. They were told that their participation was voluntary, and they were asked to sign the statement of informed consent. No subject refused to participate. After signing the consent form, the subjects were

escorted to the examination room and instructed to wait until someone came for them. The polygraph examiners were kept in another part of the building to prevent their observing what was happening to the subjects.

Guilty subjects were escorted one at a time to participate in the predetermined crime(s). Dispatch of the escorts was coordinated with walkie-talkies to prevent any subject from observing any other subject. Innocent subjects were also taken from the exam rooms for variable lengths of time to make their experience as equivalent as possible to that of the guilty subjects.

The scenarios for the three crimes were as follows. Crime 1 was the theft of a of a classified document. Subjects assigned to commit Crime 1 were escorted from the polygraph building to another building half a block away. While the escort talked with an office worker, the subject entered a walk-in vault, located a mock-classified document, and copied it on a nearby photocopier. The subject returned the document to its place in the vault and hid the photocopy on his or her person, where it remained throughout the polygraph exam.

Crime 2 consisted of photographing classified equipment. Subjects were individually escorted to another nearby building, where polaroid photographs were being made of some mock classified equipment. While the photographer took the escort into another room, ostensibly to ask some questions, the subject unobtrusively entered the room with the equipment and took a picture of it with the polaroid camera. The subject hid the picture on his or her person, where it remained throughout the polygraph test. Just after the subject and escort left the building, the photographer came running out to say that the camera had been moved and to ask if either of them had touched it. Both denied having done so.

Crime 3 was an act of sabotage. The subject was detailed to police a nearby parking lot for scraps of waste paper. The trunk of one of the cars in the lot was open, as if it were being unloaded. A box of mock classified radio tubes was visible in the trunk. A hammer was nearby. The subject smashed one of the tubes with the hammer and discarded the remnants in a trash can with the waste paper. The subject was surreptitiously observed to ensure that the crime was properly committed.

The polygraph examinations were conducted by 13 instructors from the Defense Polygraph Institute. All were polygraph examiners trained at DPI or its predecessor, all were certified by their parent organizations, and all were experienced in field polygraph work. The examiners were selected on the basis of their familiarity with the general type of tests being given and their availability. The examiners were blind to the guilt or innocence of individual subjects, but they were briefed on the details of the three mock crimes so that they could conduct the

tests realistically.

Two different types of polygraph examinations were administered. Half of the subjects were tested with three single issue examinations, and half were given one multiple issue examination. The two types of examinations differed in the nature of the pretest and in the number of issues covered on each polygraph chart.

Subjects who were given single issue tests were treated as if they were criminal suspects. That is, the examiner informed them that three crimes had been committed, and that there was reason to believe that the subject may have committed one or more of them. The subjects were given a Miranda warning, and their control questions were tailored to the individual subjects' personalities and the type of crime being covered. A stimulation (number) test was administered. Then, three single issue control question tests were conducted, one after the other. Each test covered one crime and consisted of three charts. The sequence in which the crimes were covered was systematically varied to control for possible sequence effects. The following question patterns were used in the three single issue tests. The abbreviations for question types are as follows: IR - Irrelevant, SR - Sacrifice Relevant, CQ - Control Question, RQ - Relevant Question.

Crime 1 (Theft of Document):

IR 1. Is today _____?

SR 2. Do you intend to answer truthfully each question on this test?

CQ 3. Prior to coming on active duty, did you ever steal anything from a member of your family?

RQ 4. Did you steal that classified document?

CQ 5. Prior to 1988, other than what you told me about, did you ever steal anything?

RQ 6. Do you know where that classified document is now?

CQ 7. While living in _____, did you ever steal anything from someone who trusted you?

Crime 2 (Photography):

IR 1. Is today _____?

SR 2. Do you intend to answer truthfully each question on this test?

CQ 3. Prior to coming on active duty, did you ever violate the trust of a family member?

RQ 4. Did you photograph that classified equipment?

CQ 5. Prior to 1988, other than what you told me, did you ever violate anyone's trust?

RQ 6. Do you know where any photographs of that classified equipment are now?

CQ 7. While living in _____, other than what you told me about, did you violate the trust of a friend?

Crime 3 (Sabotage):

IR 1. Is today _____?

SR 2. Do you intend to answer truthfully each question on this test?

CQ 3. Prior to coming on active duty, did you ever damage anyone's personal property?

RQ 4. Did you smash that piece of classified equipment?

CQ 5. While living in _____, did you damage anything?

RQ 6. Do you know what was used to smash that piece of classified equipment?

CQ 7. Prior to 1988, did you ever damage any public property?

Thus, there were two relevant questions and three control questions in each of the three single issue tests. The multiple issue test administered to the remaining subjects used the same six relevant questions, but used only four control questions. The question sequence for the multiple issue test was:

IR 1. Is today _____?

SR 2. Do you intend to answer truthfully each question on this test?

CQ 3. Before joining the Army, did you ever steal anything from a store?

RQ 4. Did you steal that classified document?

RQ 5. Do you know where that classified document is now?

CQ 6. Prior to 1988, did you ever steal anything?

RQ 7. Did you smash that piece of classified equipment?

RQ 8. Do you know what was used to smash that piece of classified equipment?

CQ 9. While in high school, did you ever damage anything?

RQ 10. Did you photograph that classified equipment?

RQ 11. Do you know where any photographs of that classified equipment are now?

CQ 12. Between your 13th and 18th birthday, did you ever violate the trust of another?

Regardless of the test outcome, no interrogation or additional testing was conducted. The charts were numerically scored by the examiner immediately following the test. The examiner scored respiration, skin resistance, relative blood pressure and vasomotor activity on a 7-point scale that ranged from +3 to -3. Scores were determined by comparing each physiological system at each relevant question against the greater of the two nearest control questions (one preceding, the other following the relevant question). The criteria for reactions were those taught at the Defense Polygraph Institute. Negative scores were assigned when the reaction to the relevant question was larger and positive scores were assigned when the reaction to the control question was larger. The magnitude of the score was dependent on the magnitude of the difference between the relevant and control question. The scores for each relevant question were summed across the four channels and the three charts. Scores of -3 or lower to any relevant question on a test resulted in a deceptive (DI) outcome. If the test was not deceptive, but any relevant question had a score between +2 to -2 inclusive, the outcome was inconclusive. Only if the scores on all relevant questions were +3 or higher was the test categorized as truthful (NDI).

Results

Original Examiners' Classifications

Table 18 displays the overall performance of the original examiners at the gross classification of individuals as either completely innocent or guilty to at least one crime. Decisions

Table 18. Decisions of the original examiners in Experiment 2.

Approach	Decision			
Condition	NDI	INC	DI	TOTAL
Multiple Issue Approach				
Innocent	6	3	2	11
Guilty	2	11	26	39
Single Issue Approach				
Innocent	5	6	1	12
Guilty	3	4	31	38
TOTALS	16	24	60	100

with the Multiple Issue approach on subjects who committed no crimes were 55% correct, 18% incorrect, and 27% inconclusive. Excluding inconclusives, 75% of these innocent subjects were categorized correctly. With the Multiple Issue approach subjects who committed one or more crimes were called deceptive to at least one of the crimes 67% of the time, deceptive to none of the crimes 5% of the time, and 28% were reported as inconclusive. Excluding inconclusives, 93% of the Guilty subjects were classified as deceptive to at least one of the crimes. The $\underline{X}^2$ for the multiple issue portion of Table 18 was significant, $\underline{X}^2$ (2) = 16.69, $\underline{p} < 0.01$, as was the $\underline{\text{tau}}\ \underline{C} = 0.42$, $\underline{p} < 0.001$.

Outcomes with the Single Issue approach on Innocent subjects were 42% correct, 8% incorrect, and 50% inconclusive. Excluding inconclusives, 83% of these innocent subjects were categorized correctly. With the subjects who committed one or more crimes the Single Issue approach called 82% deceptive to at least one crime, 8% deceptive to no crimes, and 10% were called inconclusive. Excluding inconclusives, 91% of the Guilty

subjects were classified as deceptive to at least one crime. The X^2 for the single issue portion of Table 18 was significant, X^2 (2) = 21.25, $p < .01$, as was the tau C = 0.54, $p < .001$.

Two Kruskal-Wallis oneway ANOVAs were conducted on these data. The first Kruskal-Wallis tested for effects of the Guilt/Innocence factor on decisions, and that effect was found to be significant, X^2 (1) = 31.52, $p < .01$. The second analysis tested for an effect of the Approach (Single, Multiple), and that analysis was not significant. Possible interactions of Guilt and Approach were tested with a parametric Guilt X Approach ANOVA. That analysis found a significant main effect for Guilt, F (1, 96) = 30.4, $p < .001$, but none of the other effects were significant.

Performance was also examined at the level of accuracy of classifications for single crimes. Since there were no significant differences in classifications for the Approach taken to testing multiple issues, this analyses was collapsed[6] across the Approach factor. Table 19 illustrates the accuracy of classification for each of the crimes with subjects who committed at least one crime. X^2 analyses were conducted on the frequency tables for the three crimes and none were significant. Overall, only 33% of the outcomes on specific individual crimes were correct. The predictive relationship for crimes 1 and 2 produced significant tau C values but they were in opposite directions, tau C = 0.28 and -0.22 respectively. These results indicate that the examinations were not able to determine which crime(s) had been committed.

Numerical Scores

Possible differences between the numerical scores of the two multiple issue approaches were tested in several ways. First, a total numerical score was calculated for each subject, and the variance in those scores was decomposed with a Guilt (Innocent, Guilty) X Approach (Single, Multiple) ANOVA. That analysis indicated that Innocent subjects (M = 25.52) produced larger total numerical scores than did Guilty subjects (M = 1.76) as shown by a significant main effect for Guilt, F (1, 96) = 30.4, $p < .001$. There were no significant effects or interactions involving the Approach factor. The positive mean numerical score for Guilty subjects is not unexpected. Subjects guilty of

---

[6]The term 'collapsed' is used to indicated that two (or more) of the original conditions of the experiment were combined for additional analysis. Collapsing across a condition is justified after a demonstration that the grouping factor being collapsed across had no statistically significant effects. In this case, since there were no significant effects for the Approach taken to multiple issue testing on the classifications obtained, it is justifiable to remove the Approach as a grouping factor from any additional analyses on classifications.

Table 19. Percent accuracy for detecting which crime was committed by subjects who committed at least one crime.

	NDI	INC	DI
Crime 1 (Espionage)			
Truthful on Crime (N = 25)	48	32	20
Deceptive on Crime (N = 52)	23	35	42
Crime 2 (Photography)			
Truthful on Crime (N = 26)	12	42	46
Deceptive on Crime (N = 51)	29	41	30
Crime 3 (Sabotage)			
Truthful on Crime (N = 26)	19	39	42
Deceptive on Crime (N = 51)	33	30	37
Combined			
Truthful on Crime (N = 77)	26	38	36
Deceptive on Crime (N = 154)	29	35	36

only one or two crimes would be expected to produce negative numerical scores to some questions and positive numerical scores to others. Those expected questions scores would combine to make the total numerical scores less extreme.

Possible differences between the crimes and between subjects based on the number of crimes they committed were tested by developing total numerical scores for the two relevant questions directed at each of the three crimes. Those crime total scores were then analyzed with a repeated measures analysis of variance (RANOVA) containing one repeated measures factor, Crime Total

Score (3 levels), and two between subject factors, Approach (Single, Multiple) and Number of crimes committed (0, 1, 2, and 3). There were no significant effects revealed by that analysis.

Discussion

The most important result in Experiment 2 was the finding of no differences in the Approach taken to testing multiple relevant issues. There were no differences between the use of one multiple issue control question test or three single issue control question tests in examiners' decisions, or in the more powerful tests of the numerical scores. Those results suggest that the multiple relevant issue testing approach was not a likely contributor to the poor detection of deception in Experiment 1. However, Experiment 2 did not examine the effect of the number of relevant questions addressed to relevant issues. In some of the question series in Experiment 1 only a single relevant question covered the acts of the scenario. The effects of the number of relevant questions devoted to the acts of deception remain to be determined.

The accuracy levels achieved in Experiment 2 are better than Experiment 1. The tau C for the two approaches averages 0.48. Since neither study offered any reward or punishment for passing or failing the examinations, this finding suggests that the lack of reward or punishment associated with examination outcomes was not the critical factor in the poor detection of deception in Experiment 1. However, the tau C obtained in Experiment 2 indicates that the examinations in this study still were not very good discriminators of guilt and innocence. The decisions in Experiment 2 only account for about a third as much variance in the Guilt/Innocence criterion as did the decisions of the Secret Service examiners in Honts et al (1988), and only about a fourth as much as a recent mock crime study (Kircher & Raskin, 1988). Further, the mean numerical score for Guilty subjects was positive, rather than negative as predicted by the rationale of the control question test. These results leave open the possibility that the lack of explicit reward or punishment associated with examination outcomes in these experiments may still be a contributor to poor detection, and they are consistent with the analysis of Kircher et al, (1988), which indicates that the motivational structure is an important variable in detection of deception experiments.

One interesting finding of Experiment 2 was that the examinations did not detect deception at the level of the individual crimes. This result has important implications for examiners who must test on multiple relevant issues, as it suggests that the numerical scores associated with individual relevant issues may be a poor guide in choosing issues for interrogation. This result suggests that when deception is inferred, the interrogator may need to address all of the relevant issues of the examination with the interrogation.

Experiment 3

Introduction

Experiment 1 differed methodologically from other detection of deception experiments in a number of ways. Almost all previous research on lie detection used relevant questions tailored specifically to the mock crime under investigation. The examinations in Experiment 1 usually used relevant questions that were worded very generally about rather broad categories of activity. The generality of screening questions could contribute to false negative errors by reducing the emotional impact or diffusing the salience of the relevant questions. Additional problems could arise with general relevant questions if the examiner did not completely define what is included in and excluded from each relevant question. If the relevant questions are somewhat ambiguous, guilty subjects might think that the relevant questions do not pertain to them and they might not respond.

Another methodological factor that differentiated Experiment 1 from previous research was the time lag between enacting the mock espionage and the running of the polygraph tests. In Experiment 1, two months elapsed between the scenarios and the polygraph tests. Some recent research (e.g., Honts, 1986; Honts, Hodes, & Raskin, 1985; Honts, Raskin, & Kircher, 1987; Podlesny & McGhee, 1987) has introduced intervals of several days or a week, but none has approached the two month time lag of Experiment 1. That amount of time could conceivably have blunted the programmed guilty subjects' emotional reaction to their scenarios.

Experiment 3 investigated the effect of having a long interval between the enactment of the mock crime and the polygraph examination, and the effect of general versus specific relevant questions.

Method

Subjects

Volunteers were initially solicited from among the 207 subjects who had served in Experiment 1. Some had been reassigned from Ft. McClellan, but 83 Experiment 1 subjects volunteered to serve in Experiment 3. An additional 17 subjects similar to those used in Experiment 2 were recruited from the basic trainees at Ft. McClellan. None of the basic trainees had ever served as research subjects. None of the 100 subjects was paid to volunteer, and no explicit reward or punishment was associated with test outcomes.

Examiners

Fifteen instructors at the Defense Polygraph Institute served as examiners. All were trained at the Defense Polygraph

Institute or its predecessor, all were federally certified examiners, and all were experienced in field polygraph work. Thirteen had served as examiners in Experiment 2. As in Experiment 2, examiners were selected on the basis of familiarity with the general type of test (screening or criminal investigation) and their availability (lack of conflict with other assigned duties). The examiners were blind to each subject's guilt or innocence, the base rate of deception, and the nature of the espionage scenario.

Apparatus

Lafayette field polygraphs were used to record respiration, cardiovascular activity, vasomotor activity, and the skin resistance response. The equipment was similar to that described for Experiment 2, except that a second respiratory channel was recorded instead of skin conductance. The examinations were conducted in May, 1988 in the same rooms used in the two earlier experiments.

Procedure

Experiment 2 was an unbalanced 2 X 2 X 2 factorial design that consisted of three between subject factors: the amount of time between enacting the crime and taking the polygraph examination (about 30 minutes versus six weeks), the specificity of the relevant questions (general versus specific), and guilt (guilty versus innocent). The 83 subjects who had participated in Experiment 1 were randomly assigned to the cells in the design matrix. However, the 17 basic trainees were available on the examination day only. Consequently they could not be assigned to the long latency condition. They were randomly assigned only to the four cells (guilty/innocent, general/specific) in the short latency condition. The design and number of subjects in each cell was as follows:

		Guilt				
		Innocent		Guilty		
		Latency				
		6 Weeks	Same Day	6 Weeks	Same Day	
Type Question	General	5	5	20	20	50
	Specific	5	5	20	20	50
		10	10	40	40	100

Subjects arrived at the Polygraph research annex either about six weeks prior to their polygraph test or on the day of the test. Each subject read a description of the study, signed a statement of consent, and then read instructions for their

assigned condition (guilty or innocent). All subjects were given a sealed envelope and instructed to take it to the Institute's main building, across the road from the annex. They went to the office of the Director's secretary to deliver the envelope. The purpose of the delivery was to give both the innocent and guilty subjects access to a mock classified document on the secretary's desk. If a polygraph examiner inadvertently saw the subject in the secretary's office, he would not know if the subject was guilty or not. All subjects used the envelope as a pretext to get the secretary out of her office, during which time the guilty subjects had to locate and steal a mock classified document from the secretary's desk.

After smuggling the stolen document out of the building, the guilty subjects read the document, then hid it in a tin can in a nearby assembly area for retrieval by another spy. Half of the subjects were tested six weeks following this initial activity and half were tested immediately.

Two types of polygraph examinations were administered. Half of the subjects were treated as if they were criminal suspects and they were given an examination with very specific relevant questions. They were informed that a classified document had been stolen from a room they had had access to. They were given a Miranda warning and they were given a pretest interview similar to that used in criminal investigative examinations. The relevant questions of the specific relevant question examination were oriented to the theft of a specific classified document. The control questions were tailored to both the crime and the subject's personality. Typical test questions for the specific relevant condition were as follows.

IR 1. Is today _____?

SR 2. Do you intend to answer truthfully each question on this test?

CQ 3. While on active duty, did you ever steal any government property?

RQ 4. Were you instructed to steal that secret document from the secretary's office?

RQ 5. Did you steal that secret document from the secretary's desk?

CQ 6. Other than what you told me, before 1988, did you ever steal anything?

RQ 7. Did you at any time read that secret message?

RQ 8. Did you hide that secret document for someone else to pick up?

CQ 9. Prior to coming to Ft. McClellan, did you ever steal anyone's personal property?

The other half of the subjects were administered a security screening type of examination with general relevant questions. They were not given a Miranda warning, and their pretest interview was similar to a counterintelligence screening examination. However, the relevant questions were not the normal counterintelligence questions. Only one or two of the questions normally included on counterintelligence screening tests would apply to the theft of a document from a secretary's desk. Consequently, if the criminal type test was found to be more accurate than the screening type, it could be due either to the specificity of the relevant questions or to the number of relevant questions dealing with the theft. To avoid that problem, both the question sequence and the number of of relevant questions that the guilty would have to lie to were held the same. The only differences were the nature of the pretest interview, the specificity of the relevant questions, and the latitude of the examiner in selecting control questions. The questions used on the general relevant question test were:

IR 1. Is today _____?

SR 2. Do you intend to answer truthfully each question on this test?

CQ 3. Have you ever deliberately done anything dishonest?

RQ 4. Have you ever planned to take classified documents without authorization?

RQ 5. Have you ever committed an act of espionage against the US?

CQ 6. Are you a really honest person?

RQ 7. Have you ever participated in providing classified information to an unauthorized person?

RQ 8. Have you ever removed classified defense material from a building without authorization?

CQ 9. Have you ever lied to make yourself look important?

Following the examination the subjects were given a debriefing questionnaire similar to the one used in Experiment 1 (see Appendix A).

Results

Original Examiners' Classifications

The overall performance of the original examiners is shown in Table 20. With Innocent subjects, the examiners' classifications were 90% correct, 10% incorrect, and none were inconclusive. With guilty subjects, the examiners' classifications were 75% correct, 7% inconclusive, and 18% incorrect. Excluding inconclusives, 81% of the Guilty subjects were classified correctly, and 19% were false negative errors.

Table 20. Decisions of the original examiners in Experiment 2.

	Decision			
Condition	NDI	INC	DI	Total
Innocent	18	0	2	20
Guilty	14	6	60	80

The predictive relationship illustrated in Table 20 was tested in the several ways described for evaluating decisions in Experiment 1. X^2 analysis was conducted, and the X^2 for Table 1 was significant, X^2 (2) = 38.68, $p < 0.01$. The tau C for the relationship illustrated in Table 1 was also significant, tau C = 0.46, $p < 0.01$.

A series of Kruskal-Wallis oneway ANOVAs was used to test for the effects of Guilt, Time Lag, and Test Type on the decisions. Only Guilt produced a significant result, X^2 (1) = 44.3, $p < 0.001$. Examiner decisions were not affected by the time lag or the specificity of the test. A Time Lag X Test Type X Guilt parametric ANOVA was also conducted on the decision data to test for the possibility of interactions between the factors, and again only the main effect for Guilt was significant, F (1, 93) = 87.2, $p < 0.001$.

Numerical Scores

Relevant Question Effects. The numerical scores were collapsed across the five physiological channels and were analyzed with a RANOVA. That analysis included three between subjects factors, Guilt (Innocent, Guilty), Time-Lag (Immediate, 6 weeks), and Test Type (Specific, General), and two repeated measures factors, Chart (3 Levels) and Relevant Question (4 Levels). The RANOVA found only two significant effects. The mean total numerical score for Innocent subjects (M = 35.4) was significantly larger than the mean total numerical score for

Guilty subjects (M = -2.3), as was indicated by a significant main effect for Guilt, F (1,92) = 69.32, p < 0.001. The other significant, but small, effect was an obscure 4-way interaction between Guilt, Time-Lag, Test Type, and Chart, F (2, 104) = 3.37, p < 0.05.

Physiological Channel Effects. The numerical scores were collapsed across the four relevant questions and were then analyzed with RANOVA. That RANOVA contained two repeated measures factors, physiological Channel (5; thoracic respiration, abdominal respiration, skin resistance, relative blood pressure, and finger pulse amplitude) and Chart (3), and three between subject factors, Guilt (Innocent, Guilty), Time-Lag (Immediate, 6 weeks), and Test Type (Specific, General). As expected, this RANOVA revealed the same main effect for Guilt and the interaction Guilt, Time-Lag, Test Type and Chart as was described above. This analysis also indicated a significant main effect for Channel, F (4, 368) = 13.01, and a significant interaction of Guilt and Channel, F (4, 368) = 7.72. The means representing these effects are shown in Table 21. The main effect for Channel

---

Table 21. Mean numerical scores of the various physiological channels by guilt condition.

Guilt	TR	AR	SRR	RBP	FPA	Combined
Innocent (n = 20)	3.8	4.6	13.4	6.2	7.5	35.4
Guilty (n = 80)	-1.0	-0.9	0.3	-0.2	-0.6	-2.3

TR = Thoracic Respiration

AR = Abdominal Respiration

SRR = Skin Resistance Response

RBP = Relative Blood Pressure

FPA = Finger Pulse Amplitude

---

appears to be primarily due to skin resistance, which produced more positive means than the other channels. The interaction of Channel and Guilt appears to be due to the various channels being more or less effective with the Innocent subjects, while they were of approximately equal effectiveness with the Guilty subjects. None of the other main effects were significant. However, one other interaction was significant. The 3-way

interaction of Guilt, Channel, and Test Type was significant, F (4, 368) = 3.87, $p < 0.05$, but is difficult to interpret.

The magnitude of the predictive validity of the physiological channels was also assessed. Scores for each of the physiological channels and their total sum were collapsed across Time Lag, Chart, Relevant Questions, and Test Type and were then correlated with the guilty/innocent criterion and with each other. The resulting correlation matrix is presented as Table 22. All correlations were significantly different from zero.

Table 22. Correlation matrix for the various physiological measures and the guilt criterion.

	TR	AR	SRR	RBP	FPA	Total Score
Guilt	-0.42	-0.48	-0.54	-0.52	-0.49	-0.65
TR		0.79	0.31	0.42	0.39	0.68
AR			0.34	0.49	0.49	0.73
SRR				0.49	0.54	0.81
RBP					0.54	0.77
FPA						0.78

All correlations are significantly different than chance.

TR = Thoracic Respiration

AR = Abdominal Respiration

SRR = Skin Resistance Response

RBP = Relative Blood Pressure

FPA = Finger Pulse Amplitude

Other than the total score, the skin resistance response produced the largest correlation with the criterion, indicating that it was the most discriminating channel, and thoracic respiration produced the smallest correlation with the criterion indicating that it was the least useful predictor.

Population Differences

Both basic training personnel and other civilian and military subjects were used in this study. Analyses were conducted to test for the possibility of differences between the trainees and the other subjects. A physiological Channel X Guilt X Time Lag X Test Type X Subject Type (Trainee, Other) RANOVA of the total numerical scores revealed no differences between the two populations, nor any interactions of Subject Type with any of the other factors.

Affective Responses

During their debriefings, subjects in both Experiments 1 and 3 gave ratings on a 10 point scale of 8 affective descriptors of their subjective responses during their polygraph examinations. Their mean responses and the associated standard deviations are presented in Table 23. The affective ratings were subjected to a RANOVA with Guilt and Study (Experiment 1, Experiment 3) as between subject factors, and one repeated measures factor, Descriptor (8 levels). The main effect of Descriptor was significant, F (7, 2074) = 51.65, $p < 0.001$, indicating that different ratings were given to different descriptors. The interactions of Descriptor and Guilt, F (7, 2074) = 8.39, $p < 0.001$, Descriptor and Study, F (7, 2074) = 7.63, $p < 0.001$, and the 3-way interaction of Descriptor, Guilt, and Study, F (7, 2074) = 5.31, $p < 0.01$, were all significant, but are not easily interpretable. Of more interest were significant main effects for Guilt, F (1, 296) = 37.28, $p < 0.001$, and Study, F (1, 296) = 7.85, $p < 0.01$, and a significant interaction of Guilt and Study, F (1, 296) = 6.13, $p < 0.01$. The sources of these between subjects effects were examined.

In order to determine which descriptors were actually different across the Guilt conditions, a series of univariate ANOVAS were conducted. The Descriptors that produced significant univariate main effects for Guilt were Nervous, F (1, 298) = 5.74, $p < 0.05$, Tense, F (1, 298) = 10.07, $p < 0.01$, Guilt, F (1, 298) = 98.0, $p < 0.001$, and Anxious, F (1, 298) = 25.6, $p < 0.001$. The means for Guilty and Innocent subjects on these descriptors are shown in Table 24.

Table 23. Means and standard deviations for the affective responses given by subjects in Experiments 1 and 3.

	Experiment 1			Experiment 3		
VARIABLE	MEAN	S.D.	N	MEAN	S.D.	N
FEAR	3.5	2.5	206	3.5	2.8	96
NERVOUS	4.8	2.8	206	5.1	3.0	96
BORED	3.5	2.8	206	2.9	2.8	96
TENSE	4.5	2.6	206	5.1	2.9	96
CURIOUS*	8.3	2.2	206	7.1	2.9	96
GUILT*	3.2	3.0	206	5.2	3.7	96
ANXIOUS	4.7	2.9	206	5.1	3.0	96
HOPE (OF NOT* BEING CAUGHT	6.8	3.5	206	5.4	3.9	96

*Significant difference between Experiment 1 and Experiment 3.

Table 24. Means and standard deviations for the affective descriptors that differed significantly across the Guilt condition.

	Innocent			Guilty		
VARIABLE	MEAN	S.D.	N	MEAN	S.D.	N
NERVOUS	4.4	2.9	132	5.2	2.7	168
TENSE	4.1	2.7	132	5.0	2.6	168
GUILT	1.9	1.9	132	5.3	3.5	168
ANXIOUS	3.9	2.8	132	5.6	2.9	168

Since there likely was a great deal of cognitive overlap between the descriptors presented to the subjects, discriminant analysis was used to determine which descriptor(s) actually

discriminated between the Innocent and Guilty conditions. Four variables loaded into the significant Discriminant solution. Tense, Guilt, and Anxious loaded as predictors with standardized discriminant function coefficients of 0.21, 0.94, and 0.39 respectively. Fear loaded into the solution as a suppresser variable with a standardized discriminant function coefficient of -0.49. The coefficients indicate that most of the discrimination between Innocent and Guilty subjects was carried by Guilt. The fact that Nervous dropped out of the discriminant analysis indicated that Nervous was completely redundant with the retained variables.

The main effect of Study (Experiment 1, Experiment 3) was decomposed in a similar manner. The Descriptors were subjected to a series of univariate analyses, and three were found to be significant: Curious, F (1, 298) = 14.08, $p < 0.001$, Guilt, F (1, 298) = 28.11, $p < 0.001$, and Hope, F (1, 298) = 7.56, $p < 0.01$. The means for these variables are shown in Table 23. A discriminant analysis was conducted on the descriptor ratings with Study as the criterion. Curious, Guilt, and Hope contributed significantly to the discrimination, with standardized discriminant function coefficients of, 0.48, -0.84, and 0.38, respectively. Again the Guilt variable accounted for most of the discrimination.

To decompose the interaction of Guilt and Study, univariate Guilt X Study ANOVAs were conducted on each of the Descriptors. Only the analyses of Guilt and Hope produced significant Guilt X Study interactions, F (1, 296) = 5.94, $p < 0.05$, and F (1, 296) = 23.11, $p < 0.01$, respectively. The means for these two interactions are shown in Table 25. Innocent subjects in Experiment 3 reported feeling less guilt than Innocent subjects in Experiment 1, while Guilty subjects in Experiment 3 reported feeling more guilt than Guilty Subjects in Experiment 1. Innocent subjects in Experiment 3 gave smaller ratings on the Hope descriptor than did Innocent subjects in Experiment 1.

Table 25. Mean responses to the affective descriptors of Guilt and Hope, across Guilt and Study.

	Experiment 1	Experiment 3
Affective Descriptor		
Condition		
Guilt		
Innocent	2.02 (116)	1.38 (16)
Guilty	4.61 (92)	6.12 (76)
Hope		
Innocent	6.86 (116)	1.31 (16)
Guilty	6.59 (92)	6.32 (76)

Discussion

There are two major findings from Experiment 3. The first of those findings is that the time lag between the mock espionage and the polygraph examination had no effect on either the examiners' decisions or on the numerical scores. This result indicates that the time lag used in Experiment 1 probably did not contribute to the poor detection of deception. Interestingly, it also suggests that the inclusion of a time lag in analog studies of the detection of deception is probably not necessary for generalization. This is an important methodological finding that is supported by research at the University of Utah (Honts et al, 1985; 1986; 1987; Horowitz, Raskin, Honts, & Kircher, 1988).

The second important finding of Experiment 3 is that the specificity of the relevant questions had no effect on either decisions or numerical scores. This result suggests that the use of relevant questions with general wording in Experiment 1 probably did not contribute to the poor rates of detection of deception.

This study may also provide some insight into the question of motivation. As in Experiment 2, the examinations used in this experiment were better discriminators of truth and deception than were the examinations given in Experiment 1. The obtained tau C of 0.46 in Experiment 3 indicates that the decisions in Experiment 3 accounted for about twice the variance in the guilt/innocent criterion as did the decisions in Experiment 1, but only about a third as much variance as did the examiners' decisions in Raskin et al. (1988), and about a fourth as much as variance as did the blind evaluator in a recent mock crime study (Kircher & Raskin, 1988). Further, the mean numerical score from the guilty subjects in Experiment 3 was only -2.3. This result is closer to zero than would be predicted from either the rationale of the control question test, or from most of the analog detection of deception literature. There are a number of factors that might account for that result, on of which is the lack of explicit reward or punishment associated with the relevant questions. It is conceivable that the lack of motivation associated with the polygraph examinations' outcome could have effected the results of all of the studies in this report.

Experiment 3 found that the GSR was the most useful physiological measure. The GSR has been shown to be the most useful physiological measure in virtually every published study of the detection of deception, yet to date no major numerical scoring system has been altered to explicitly take advantage of this information. Research is underway at the Defense Polygraph Institute exploring ways of modifying the numerical scoring system to take optimal advantage of the GSR.

An interesting methodological finding of Experiment 3 was that there was no difference in the accuracy of polygraph

examinations given to support troops and to other personnel at Fort McClellan. This is an important finding because it suggests that support troops are an acceptable subject pool for use in detection of deception research.

The final topic for discussion in Experiment 3 concerns the affective descriptors endorsed by subjects in Experiments 1 & 3. In general, the subjects in both experiments did not strongly endorse the negative descriptors of fear, nervous, tense, anxious, guilt, or bored. They did strongly endorse the descriptor curious. These results suggest that the affective environment induced in these analog studies was not very similar to that in the field. To the extent that these studies did not re-create the environment of the real world their generalizability may be limited. We will return to the issue of generalizability in the next section of this report.

There were some significant differences in the affective descriptors endorsed by the subjects in Experiments 1 & 3. The subjects in Experiment 3 reported less curiosity and hope but more guilt than the subjects in Experiment 1. These results are difficult to interpret, but suggest that the subjects found Experiment 3 to be relatively more negative than Experiment 1. Similarly, the interactions of Guilt and Study for the descriptors Hope and Guilt are difficult to interpret, but generally seem to indicate that the subjects found Experiment 3 to be a more negative experience.

GENERAL DISCUSSION

The research conducted as Experiments 2 and 3 suggests that the three methodological issues raised in the discussion of Experiment 1 cannot adequately explain the poor detection in Experiment 1. Experiment 2 failed to find any problem specifically associated with the examination of multiple relevant issues within one question series. Experiment 3 failed to indicate any effect of a time lag between the mock espionage and the polygraph examination. Finally, Experiment 3 failed to find any problem with using relevant questions that are worded generally and presented in a screening examination as compared to relevant questions with very specific wording presented in a criminal investigative examination.

The lack of explicit rewards or punishments associated with the outcomes of the examinations may make it easier for both guilty and innocent subjects to pass the test. In Experiment 1, subjects were told that admissions to any real world security violations would be adjudicated. Those instruction may have increased the power of the control questions, possibly to the point of overwhelming relevant questions about the programmed scenarios. Those instructions represent a confounding factor in the results of Experiment 1. The effects of motivation on the detection of deception need to be systematically examined in future research. Thus, the studies reported here may have overestimated the number of false negative errors and underestimated the number of false positive errors in the field. Since these uncertainties about the effects of the laboratory remain strong generalization of the results of these studies to the field is not possible. However, despite those uncertainties it seems likely that the these studies accurately reflect trends in the real world.

One way to estimate the generalizability of the results of experiments is to use real world outcome rates and a conditional probability analysis to map the experimental outcomes on to a real world data set. The Department of Defense Polygraph Program Report to Congress for Fiscal Year 1986 and the Department of Defense Polygraph Program Report to Congress for Fiscal Year 1987 provide one such data base. During fiscal years 86 and 87, DoD components conducted 8599 security screening examinations under the congressional test program. Of those 8599 examinations, no opinion was rendered on 11 cases, 7 were reported as inconclusive, 8528 were reported as no deception indicated, and 53 were reported as deception indicated. All of the cases reported as deception indicated were confirmed by confession. Most of the reported confessions were to acts classified as security violations, rather than to espionage. These data can be used as a base for a conditional probability analysis.

However, the DoD reports make no estimate of the base rate of deception. We decided that a rough estimate of the base rate of security violation targets could be obtained from Experiment 1.

General Discussion

In Experiment 1 the three agencies that stressed gaining admissions obtained a real world admission rate of about 20%. This is likely to be a conservative estimate of the actual base rate for security violations, since it represents only those individuals who actually admitted violations. Almost certainly there were additional individuals who had committed security violations but did not admit them. However, an accurate estimate of the actual base rate is not available, and for purposes of this discussion we decided to use 20% as our base rate of deception for a conditional probability analysis of the data from the DoD reports[7]. Additionally, in order to simplify the analysis we are ignoring no opinion and inconclusive outcomes.

A conditional probability analysis using the overall results from Experiment 1 produces results somewhat similar to those predicted by the studies of criminal investigative examinations[8]. That is, our conditional probability analysis predicted that 86% of the NDI outcomes should be correct, but only 74% of the DI outcomes were predicted to be correct. In other words, there should be a large number of false positive errors. However, the DoD reports do not provide support for this analysis. The above analysis predicts 789 DI outcomes, but only 53 DI outcomes were reported. This result suggests that the overall estimates of accuracy obtained from Experiment 1 are not an accurate reflection of screening in the congressional test programs.

However, a closer examination of the results of Experiment 1 suggests that the Agency 4 examiners were performing most like the examiners who's results are reported by DoD. When we performed a conditional probability analysis using the Agency 4 accuracy rates from Experiment 1 on the population of 8581 DoD examinations, and used a base rate of 20%, we predicted no false positive errors and expected to correctly detect 137 guilty individuals[9]. The DOD reports found no false positive errors and reported 53 people as deceptive. Of course, the important implication from this analysis is that it suggests that there were more than a 1500 individuals who committed security violations, but were cleared by the polygraph (however, see Footnote 7).

---

[7]The following points should be considered in evaluating the two conditional probability analyses that follow. First, our assumption of a 20% base rate of deception is not reasonable for agencies that are not concerned with detecting security violations. Agency 4's screening program is directed at acts of espionage and sabotage and does not recognize security violations as within the scope of their polygraph program, as defined under DoD Directive 5210.48. Since Agency 4 does not even record the security violations that are reported by their subjects, the base rate of targeted deception for their programs must be considered to be very low. Second, these conditional probability analyses are based on real world data obtained from the congressional test program within the Department of Defense. Therefore, the results may not apply to programs not included in the DoD reports to congress.

General Discussion

If this analysis is correct, then it is important to consider why the results of screening tests are so biased toward NDI calls. A possible line of reasoning is not difficult to develop. The general tone of the scientific testimony before congressional committees has been that there inevitably must be a large number of false positive errors in mass screening, and this must be so unless the discriminator is nearly perfect with innocent individuals. Knowing the dire predictions of large numbers of false positive errors, it is possible that the individuals who set up the extant screening programs built in as many safeguards as possible against making false positive errors. However, they may have gone too far in protecting against false positive errors. We may have a system that is efficient at avoiding false positive errors at the expense of missing the targets it was designed to catch.

Discussions with experienced screening examiners indicates that there are a variety of pressures acting upon the examiners to clear as many examinees as possible. The primary pressure appears to stem from the knowledge that the proportion of actual espionage agents within the population being tested is extremely small. It is no wonder that if the person taking the test is having trouble clearing it, many examiners feel that they (the examiners) must be doing something wrong. We have been told that sometimes examiners run repeated tests until the physiological

---

[8]Conditional probability analysis assumptions:

Population of Cases Where a Decision was rendered: 8581

Base Rate of Guilt: 20%, therefore: Number of Guilty = 1716
and Number of Innocent = 6865

Accuracy rates from the combination of the agencies from Experiment 1: Guilty = 34% Correct, Innocent = 97% Correct

Predicted Classification Table

	DI	NDI	Totals
Guilty	583	1133	1716
Innocent	206	6659	6865
Totals	789	7792	8581

Confidence in a DI outcome = 74% (583/789)
Confidence in a NDI outcome = 86% (4228/4970)

reactions disappear without any significant admissions having been made. Although polygraph program managers try to combat that attitude, it is encouraged by other aspects of the system. Examiners who clear more subjects than the average examiner are often promoted more rapidly. Conversely, examiners who consistently fail to clear enough subjects on the first series are "given help". In such a testing environment, the results of Experiment 1 may reflect the condition of the real world. Given the astonishingly small number of positive outcomes reported in the DoD program, it seems likely that the DoD screening programs are missing a lot of security problems.

The screening situation suggested by the results of these studies and analyses is not good, but some positive aspects were indicated. The government does obtain the benefit of uncovering some security problems. A detection rate of 34% was demonstrated across the agencies. Some utility of the polygraph test was demonstrated for several of the Agencies in Experiment 1 by the substantial number of real world security problems that were discovered. Without the polygraph, it is likely that none of the problems would have been uncovered. The ability to detect some problems is better than detecting none at all. Furthermore, any possibility of being caught may deter potential spies and reduce security violations.

---

[9]Conditional probability analysis II assumptions:

Population of Cases Where A Decision was rendered: 8581

Base Rate of Guilt: 20%, therefore: Number of Guilty = 1716
and Number of Innocent = 6865

Accuracy rates from Agency 4 in Experiment 1:
Guilty = 8% Correct, Innocent = 100% Correct

Predicted Classification Table

	DI	NDI	Totals
Guilty	137	1579	1716
Innocent	0	6865	6865
Totals	137	8444	8581

Confidence in a DI outcome = 100% (137/137)
Confidence in a NDI outcome = 81% (6865/8444)

General Discussion

The security screening problem is difficult, and may be very difficult to solve. The situation could be improved by improving our ability to detect deception, and by making better use of all of the available information. However, given the current state of the practice in the field that may be very difficult to accomplish. At present, at least four questioning techniques are used and none have received sufficient scientific evaluation. Further, we were told that in at least one agency chart evaluation varies from office to office, and perhaps from one quality control officer to another. Different agencies have very different perceptions about how examinations should be conducted and about what are the appropriate targets of their screening programs. Standardization driven by research is needed.

There are a number of approaches that research could offer to improve the situation. Statistical approaches to decision making would surely help reduce the unreliability in the current systems. Discriminant analysis procedures that make explicit use of base rate information are one step that could provide an immediate benefit, and they are currently available. New approaches to analysis of physiological responses also hold promise. For example, actuarial decisions can be made on the basis of the pattern of physiological responses to relevant questions, and those decisions were demonstrated to be more accurate than decisions based on numerical scoring in one study Honts, Kircher, and Raskin (1988). New physiological measures may improve our ability to detect deception. However, all of these avenues require research, and the need is urgent if the results of Experiment 1 tell us anything about the current performance of counterintelligence screening examinations.

REFERENCES

Barland, G. H. (1981). A Validity and Reliability Study of Counterintelligence Screening Test. Report prepared for Security Support Battalion, 902d Military Intelligence Group, Ft. George G. Meade, MD. Available from the author.

Barland, G. H. (1988) The polygraph test in the US and elsewhere. In A. Gale, (Ed.), The Polygraph Test: Lies, Truth and Science. Beverly Hills, CA: SAGE.

Correa, E. J. & Adams, H. E. (1981) The validity of the pre-employment polygraph examination and the effects of motivation. Polygraph, 10, 143-155.

Department of Defense (1986). Department of Defense Polygraph Program: Report to Congress for Fiscal Year 1986. Washington, D.C.: Department of Defense.

Department of Defense (1987). Department of Defense Polygraph Program: Report to Congress for Fiscal Year 1987. Washington, D. C.: Department of Defense.

Honts, C. R. (1986). Countermeasures and the physiological detection of deception: A psychophysiological analysis. Dissertation Abstracts International, 47, 1761B. (Order No. DA8616081)

Honts, C. R., Hodes, R. L., & Raskin, D. C. (1985). Effects of physical countermeasures on the physiological detection of deception. Journal of Applied Psychology, 70, 177-187.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1986, August). Countermeasures and the detection of deception. Paper presented at the annual meeting of the American Psychological Association, Washington, D. C.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1987). Physical countermeasures may reduce the physiological detection of deception. Journal of Psychophysiology, 1, 241-247.

Honts, C. R., Raskin, D. C., Kircher, J. C., & Horowitz, S. W. (1988, March). A field validity study of the control question test. Paper presented at the American Psychology and Law Society / Division 41 Midyear Conference, Miami, Florida.

References

Horowitz, S. W., Raskin, D. C., Honts, C. R., & Kircher, J. C. (1988, October). Control questions in physiological detection of deception. Paper presented at the annual meeting of the Society for Psychophysiological Research, San Francisco, CA.

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

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

Lykken, D. T. (1981) A Tremor in the Blood: Uses and Abuses of the Lie Detector. New York: McGraw-Hill.

Norusis, M. J. (1986). SPSS/PC+ for the IBM PC/XT/AT. Chicago: SPSS Inc.

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

Podlesny, J. A., & McGhee, C. M. (1987). Investigative detection of deception: Role discrimination with a general question technique. Psychophysiology, 24, 605-606.

Raskin, D. C. (1984) Statement submitted to Committee on Armed Services, United States Senate, 7 March.

Raskin, D. C. (1986a) The polygraph in 1986: Scientific, professional and legal issues surrounding application and acceptance of polygraph evidence. Utah Law Review, 1, 29-74.

Raskin, D. C. (1986b) Testimony before the committee on Labor and Human Resources, 23 April 1986, United States Senate (Hearing on S 1815, Polygraph Protection Act of 1985). Washington, D. C. : Government Printing Office.

Raskin, D. C. (1988). Methodological issues in estimating polygraph accuracy in field applications. Canadian Journal of Behavioral Science, 19, 389-404.

Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill.

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). A Study of the validity of polygraph examinations in criminal investigation. (Final report Grant No. 85-IJ-CX-0040). Washington, D.C.: National Institute of Justice.

GLOSSARY

ANOVA -- Analysis of variance. A powerful statistical technique used in complex factorial experiments to determine if the contribution of the various factors (independent variables) and their combinations (interactions) to the total variation in the results is significantly different than that expected by chance. Also see t-test and RANOVA.

Cardio -- Short for cardiosphygmograph, one of the three channels usually recorded by field polygraph instruments. It provides a measure of relative blood pressure by measuring changes in the volume of the upper arm (sometimes the lower) by means of a pressure cuff.

Chi square ($\underline{X}^2$) --A type of statistical test (named after the Greek letter chi) used in this study to determine if the number of subjects in the various outcomes were distributed by chance.

CIA -- Central Intelligence Agency.

Correct Rejection -- In lie detection, diagnosing an 'Innocent' person as not deceptive (NDI or NSR).

DI -- Deception Indicated. A polygraph outcome in which the examiner concludes that the person is deceptive or concealing information. It is synonymous with SPR (specific reaction), and the opposite of NDI (no deception indicated) and NSR (no specific reaction).

False negative -- A polygraph outcome in which a deceptive ('guilty') person is erroneously diagnosed as truthful by the examiner.

False positive -- A polygraph outcome in which a truthful ('innocent') person is erroneously diagnosed as deceptive by the examiner.

FN -- See false negative error.

FP -- See false positive error.

GSR -- Galvanic skin response. One of the three physiological measures usually recorded by field polygraph instruments. It represents an emotional sweating response. The specific measurement of GSR used in this study is the skin resistance response (SRR).

Glossary

Guilty -- In this report, a guilty person was one who had committed a mock crime, such as the theft of a mock classified document, and they were instructed to lie about their involvement on the polygraph. If the examiner concludes he is deceptive, the result is a 'hit' (true positive). If the examiner clears him, the result is a 'miss' (false negative). The the outcome is inconclusive, it is IG (an inconclusive outcome on a guilty person).

Hit -- In lie detection, calling a 'Guilty' person deceptive (DI or SPR).

IG -- An inconclusive outcome on a person programmed to be guilty or knowledgeable.

II -- An inconclusive outcome on a person programmed to be innocent.

Inconclusive -- The outcome of a polygraph examination when the examiner is unable to make a decision about a person's truthfulness. It is usually not considered to be an error. However, in screening situations the practical result is similar to a DI outcome, in that further investigation is required and clearance may be withheld pending resolution.

Innocent -- In this report, a person who was not programmed to be guilty or knowledgeable. Some programmed innocent persons may in fact not be innocent if they concealed significant real-life information from the polygraph examiner. Generally, a truthful outcome with and 'Innocent' subject is considered to be a correct decision (true negative, correct rejection). While a deceptive outcome with an 'Innocent' is generally considered to be an incorrect decision (false positive, false alarm). However, if, following a deceptive outcome, the programmed innocent person admits to concealing real-life information, that outcome may be considered to be a correct decision (true positive, hit).

INSCOM -- US Army Intelligence and Security Command. See MI.

Knowledgeable -- In this report, a person who is programmed to have knowledge of someone who has committed a mock crime, but who did not commit the crime himself. If the examiner concludes that the person is lying or concealing information on the test the outcome is considered to be a correct decision (hit, true positive). If the outcome was NDI it was considered to be an incorrect decision (miss, false negative).

MI -- Military Intelligence. In this report, MI refers specifically to the US Army Intelligence and Security Command (INSCOM) and its subordinate elements.

Miss -- An error of diagnosing a 'Guilty' subject as truthful (NDI or NSR).

NDI -- No Deception Indicated. A polygraph outcome in which the examiner concludes that the person was truthful, and was not holding back any significant information. Synonymous with NSR. The opposite of DI and SPR.

NSA -- National Security Agency.

NSR -- No specific reaction. NSA examiners use this term in preference to NDI, to indicate that a person appeared truthful on the polygraph test.

OSI -- The U.S. Air Force Office of Special Investigations.

p < -- statistical notation for 'The probability that this result could have occurred purely by chance is less than...' In this study, results which have probabilities of .05 or less are assumed to have been caused by the factor being studied, rather than by chance. See probability.

Probability -- Probability is a statistic expressed as a number ranging from .0 to 1.0, in which the smaller the number, the less likely the event. A probability of .05 means that there is only five chances out of a hundred (one in twenty) that a given result could have occurred.

RANOVA — Repeated Measures Analysis of Variance. A special case of ANOVA where one or more of the dependent variables is a repeated measure from the same subject. An example would be, the charts of a polygraph test. RANOVA takes statistical advantage of the fact that the repeated measures from the same individual are not independent observations.

SPR -- Specific Reaction. NSA examiners use this term in preference to DI, to indicate that a person appeared deceptive on the polygraph test.

tau c -- A non-parametic measure of association. For practical purposes tau c values can be treated as correlation coefficients.

TN -- See True negative.

TP -- See True positive.

True negative -- When a programmed innocent person is called truthful (NDI) on the polygraph. Synonymous with correct rejection.

True positive -- When a programmed guilty or knowledgeable person is called deceptive on the polygraph.

t-test -- A powerful statistical test used to determine if the effects of a single independent variable that has two levels are significantly different than that expected by chance. Mathematically, t-tests are a special case of ANOVA.

APPENDIX A
EXPERIMENT 1 FORMS

CONSENT FOR POLYGRAPH EXAMINATION

I, _____, voluntarily consent to polygraph testing administered by examiners of the United States Government. I understand that polygraph testing and periodic retesting can be required as a condition of my employment with the United States Military.

The procedures that are to be followed during the examination have been explained to me, and I am aware that the procedures will include the use of sensors to record my physiological responses to questions. I understand that the questions to be asked during the examination will be only those questions necessary to resolve security and counterintelligence issues, including but not limited to specific issues such as loyalty, the compromise of classified information, and vulnerability to blackmail, and that the questions will be reviewed with me, at least in general, prior to the examination. I agree to keep the details of the examination secret from all unauthorized persons.

I understand that any information relating to violations of law or an imminent threat to life or property may be reported to the Attorney General as required by Section 535 of Title 28 of the United States Code and Executive Order 12333 or its successors, and also may be reported to appropriate law enforcement or other government agencies for administrative, investigative or legal action. I also understand that I have a right against self-incrimination under the Fifth Amendment to the Constitution of the United States and that I may refuse to answer a question if my answer would tend to incriminate me.

I also have been briefed that any active duty member of the United States Armed Forces must be advised during the initial pretest, prior to signing this consent form, that any violation of Article 31/U.C.M.J. might be reported to their respective military service.

I understand the session with the polygraph examiner may be monitored and is audio and video recorded for the purpose of clarity and accuracy. I also understand that the session may be videotaped for the purpose of research and training.

I have read the foregoing and understand its import fully.

IN WITNESS WHEREOF, I place my signature below, this _____ day of _____ 19___.

_____

The above was read and signed in my presence this _____ day of _____ 19___.

_____

Appendix A

PERSONAL DATA FORM

THIS FORM IS AFFECTED BY THE PRIVACY ACT OF 1974

1. AUTHORITY: 10 USC 3012, 44 USC 3101 and 10 USC 1071-1087

2. PRINCIPAL PURPOSE: To plan involvement in a classified scenario.

3. ROUTINE USES: The requested information will be used to tailor the details of a classified scenario to those individuals selected for participation. This form will be destroyed (a) in the event you are not selected for a scenario or (b) following your participation in the scenario. None of the information will be furnished to anyone not directly involved in the research.

4. MANDATORY OR VOLUNTARY DISCLOSURE: Disclosure is voluntary. Failure to provide the information may result in your being disqualified for participation in the study.

PLEASE PRINT ALL INFORMATION AS LEGIBLY AS POSSIBLE.

Name: _____ Age:_____ Sex: _____

Height: _____ Weight: _____Race: White Black Other_____

POB:_____ Duty title: _____

Residence address: _____

Residence phone: _____ Marital status: Single Married Other

Do you have (or have access to) a vehicle? _____ Type: _____

Make: _____Color: _____

PART B — VOLUNTEER AFFIDAVIT

I, _____, being at least 18 years old, do hereby volunteer to participate in a research study entitled "Polygraph Screening Validation Study" being conducted at the Department of Defense Polygraph Institute at Ft. McClellan under the direction of Gordon H. Barland, Ph.D.

The implications of my participation; the nature, duration and purpose, and the methods by which it is to be conducted; and the inconveniences and hazards to be expected have been thoroughly explained to me as described above. I have been given the opportunity to ask questions concerning this study, and any such questions have been answered to my satisfaction. Should any further questions arise concerning my rights on study-related injury, I may contact COL Cadol, M.D., Director of the Noble Army Community Hospital, Ft. McClellan, Alabama, 36205 (Telephone number: 205/238-2200).

I understand that I may at any time revoke my consent and withdraw from the study without prejudice.

_____ _____

Signature Date

_____

Witness

Appendix A

Code number: _____

Date: _____

POLYGRAPH ATTITUDES QUESTIONNAIRE

INSTRUCTIONS: Read each sentence until you understand what is being asked. Circle the answer which best describes your attitude.

1. The polygraph, or "lie detector", is _____ able to tell when a person is lying.

   a. always b. usually c. sometimes d. rarely e. never

2. If I were suspected of a crime which I had actually committed, I would _____ agree to take a polygraph test.

   a. definitely b. probably c. might d. probably not e. never

3. If I were suspected of a crime which I had not committed, I would _____ agree to take a polygraph test.

   a. definitely b. probably c. might d. probably not e. never

4. If I were considered for a government job involving access to secret information and were asked to take a preclearance polygraph test on my background, I would _____ agree.

   a. definitely b. probably c. might d. probably not e. never

5. If I were being considered for a job in a supermarket involving access to money and were asked to take a preemployment polygraph test on my background, I would _____ agree.

   a. definitely b. probably c. might d. probably not e. never

6. Use of the polygraph _____ violates a person's privacy.

   a. never b. rarely c. often d. usually e. always

7. Use of the polygraph is _____ unethical.

   a. never b. rarely c. often d. usually e. always

8. Comments: _____

_____

Appendix A

Code number _____

Date: _____

POLYGRAPH ACCURACY QUESTIONNAIRE

INSTRUCTIONS: Answer the following questions with the percentage which best describes how you feel. Don't worry if your answers are not consistent. Few people are consistent about something like the lie detector. We are interested in your initial reaction to the question. Answer the questions as rapidly as feasible.

1. How accurate do you think the polygraph is in general? _____%

In a murder case? _____%

With the guilty person? _____%

With an innocent person? _____%

In preemployment screening? _____%

With someone who's lying? _____%

With someone who's telling the truth? _____%

When a person is lying about which of 5 numbers he picked? _____%

2. How accurate do you think the polygraph would be on you? _____%

Appendix A

S _____

Date: _____

Agency: _____

1987 SCREENING STUDY

S DEBRIEFING FORM

INTRO: Your participation in the study is now over. You are free to talk to me. Must be absolute truth, despite possible prior instructions to contrary. Do not discuss w/ friends your role or your exam until 15 Sep 87.

S first name: _____

1. How did you like your exam? _____ Was it what you expected? _____ What was different?

2. What was the best thing about it — the most interesting thing?

3. What was the worst thing about it?

4. Were Y in the Guilty group, the Knowledge group, or the Innocent group? (Check w/ our records). G_____ K_____ I_____

SCENARIOS (GUILTY and KNOWLEDGEABLE Ss):

5. How did you enjoy your scenario?

6. Was it realistic?

7. What was the most realistic thing about it?

8. What was the least realistic thing about it?

9. How could it have been improved?

10. DYK anyone who was Guilty/Knowledgeable? Who? Circumstances:

10a. Did you tell anyone about what you did, prior to the polygraph test?

11. Who else knows what you did?

12. Does anyone (else) suspect that you were guilty (or knew someone who was)?

S Debriefing

GUILTY

13. What did you do with the money?

Did holding the money for all this time cause you any problems?

15. When can we arrange to retrieve it?

16. DYH any other equipment that you've not yet turned in?.

POLYGRAPH (ALL):

17. How interesting did you find your polygraph experience?

18. What was the most interesting thing about it?

19. What was the least pleasant thing about it?

20. If you had the power and the authority to make any change in the polygraph test that you wanted to, what would be the first thing that you would change?

21. DY lie to any of the Qs? Which one(s)? Did any of the questions trouble you?

21a. Was your polygraph test accurate?

22. What does the examiner look for when he's deciding whether you were a spy or not? How do you suppose the test is graded?

23. While the test was in progress, did you feel yourself react to any of the questions? Which ones?

24. What did you do in order to look as truthful as possible on the test?

25. DY control your breathing? _____ How?

26. What did you think about while you were attached to the polygraph and the questions were being asked? _____

26a. Did you try to keep calm? _____ On just some of the questions?

26b. Did you try to look guilty on any of the questions?

26c. Did you try to create reactions to any of the questions?

Which ones?

S Debriefing

How?

27. Did any question on the test take you by surprise or catch you off-guard?

27a. What was your reaction?

27b. Why do you suppose they were put there?

28. To what extent do the followng words describe how you felt during the actual test? (Scale of 1-10)

Fear _____ Of what?
Nervousness _____
Boredom _____
Tenseness _____
Curiosity _____
Guilt _____
Anxiety _____
Hope (of not being caught?) _____

29. What single word best describes how you felt on the actual test?

30. GUILTY/KNOWLEDGEABLE: Were you hoping to beat the test ____, or were you hoping that your lies would be detected _____?

31. To what extent did you feel your polygraph examination was "for real?" (1-10) _____

32. To what extent did you feel it was just a game? (1-10) _____

33. Were you mistreated in any way by the examiner?

34. Would you be willing to volunteer for another polygraph test on the next research study we do? _____

35. Is there anything else you'd like to mention?

NOTE: Have S fill out 3 Questionnaires: Test Program, pg accuracy, and pg attitudes.

A couple of months from now we need to have you fill out these three questionnaires one more time in order to see whether any changes that occurred are short-term or long-term changes. These forms will be mailed to you. What address will you be at two months from now?

APPENDIX B
AGENCY EVALUATION QUESTIONNAIRES

BLIND EVALUATION Date _____

_____ SUBJECT CODE NUMBER _____ ETYPE _____ Evaluator

PLEASE READ THIS ENTIRE FORM BEFORE YOU VIEW ANY EXAMINATION TAPES!

We want you to give us as objective an opinion as possible about the examination you are going to watch. We realize that some of the judgments we are asking you to make are difficult ones, and that this is a tedious task. However, your job is a very important one in helping us to understand the results of this study. Please do the best you can, and please give us a response to every rating scale item, even if you are not as sure as you would like to be about your response.

Please watch the pretest interview for this examination. Take your time and observe the examination carefully. Take notes in the space provided. If you need additional space for notes please use the backs of these pages or attach additional sheets. We are particularly interested in any significant differences between this examination and the way your agency conducts examination in the field. After you have finished watching the pretest interview, and before you watch the remainder of the examination, answer the questions following the notes section.

Pretest interview notes _____

_____

_____

_____

_____

_____

Questions: Please indicate the response that most closely expresses your opinion by circling the number and writing the number in the blank at the left.

_____ In general, how much was this pretest like a pretest conducted by your agency in the field?

1 2 3 4 5 6 7
Not at all like the field Just like the field

_____ With regard to the length of the pretest, was this pretest longer, shorter, or the same length as a typical pretest conducted by your agency in the field?

1 2 3 4 5 6 7
Shorter About the Same Longer

Appendix B

_____ Was the examiner's description of the polygraph and the physiology of the detection of deception typical of those given by your agency in field examinations?

1 2 3 4 5 6 7
Less than the field About the Same More than the field

_____ Were admonitions about movement and/or breathing about the same as, stronger, or weaker than those given in the field?

1 2 3 4 5 6 7
Weaker Than the Field Same as the Field Stronger Than the Field

_____ Was the presentation and definition of the **RELEVANT** questions similar or dissimilar to the definition and presentation used by your agency in the field?

1 2 3 4 5 6 7
Not at All Like The Field Just like the field

If you felt that the definition and presentation of the **RELEVANT** questions was different from that used by your agency in the field please tell us about those differences.

_____ Was the emphasis placed on the **RELEVANT** questions in this pretest typical, less, or more than the emphasis placed on relevant questions during field examinations conducted by your agency?

1 2 3 4 5 6 7
Less emphasis About the Same More Emphasis

If you felt that the amount of emphasis placed on the **RELEVANT** questions was different from that used by your agency in field examinations, please tell us about the differences.

_____ Was the presentation and definition of the CONTROL questions similar or dissimilar to the definition and presentation used by your agency in the field?

1 2 3 4 5 6 7
Not at all like the field Just like the field

If you felt that the definition and presentation of the CONTROL questions was very different from that used by your agency in the field, please tell us about those differences.

_____ Was the emphasis placed on the CONTROL questions in this pretest typical, less, or more than the emphasis placed on relevant questions during field examinations conducted by your agency?

1 2 3 4 5 6 7
Less emphasis About the Same More Emphasis

If you felt that the amount of emphasis placed on the CONTROL questions was very different from that used by your agency in field examinations. please tell us about the differences.

_____ Based on your observation of this pretest interview, if the subject was actually GUILTY do you think this pretest interview would produce an accurate or inaccurate outcome?

1 2 3 4 5 6 7
Very Inaccurate Very Accurate

Briefly, why do you feel the way you do?

Appendix B

---------- Based on your observation of this pretest interview, if the subject was actually **INNOCENT** do you think this pretest interview would produce an accurate or inaccurate outcome?

1 2 3 4 5 6 7
Very Inaccurate Very Accurate

Briefly, why do you feel the way you do?

---------- Given your observation of the subject's behavior and statements, do you believe the subject to be guilty or innocent? (If you think the subject has guilty knowledge consider him/her to be guilty.)

1 2 3 4 5 6 7
Guilty Innocent

Please give us any general comments you may have about this pretest examination.

Appendix B

Now please watch the intervals between the charts, and take notes if you wish.

Please answer the following questions by indicating the response that most closely expresses you opinion by circling the number and writing the number in the blank at the left.

_____ In the between chart interval, were the **RELEVANT** questions discussed, and if they were, did the discussion emphasize the **RELEVANT** questions more, less, or about the same as they would be in a typical examination conducted by your agency?

1	2	3	4	5	6	7
Not discussed	Less emphasis		About the same			More emphasis

_____ If you responded 'Not discussed' is that standard practice for your agency? YES NO

_____ In the between chart interval, were the **CONTROL** questions discussed, and if they were, did the discussion emphasize the **CONTROL** questions more, less, or about the same as they would be in a typical examination conducted by your agency?

1	2	3	4	5	6	7
Not discussed	Less emphasis		About the same			More emphasis

_____ If you responded 'Not discussed' is that standard practice for your agency? YES NO

Please let us have any additional comments you may have on this examination.

NONBLIND EVALUATION Date _____

_____ SUBJECT CODE NUMBER _____ ETYPE _____ Evaluator

Appendix B

_____ Examination outcome Correct/Incorrect _____ Scenario

This subject produced a polygraph outcome as indicated. We want you to give us as objective an opinion as is possible about this examination, and why its outcome was correct or incorrect. We realize that some of the judgments we are asking you to make are difficult ones, and that this is a tedious task. However, your job is a very important one in helping us to understand the results of this study. Please do the best you can, and please give us a response to every rating scale item, even if you are not as sure as you would like to be about your response.

Please watch the pretest interview for this examination. Take your time and observe the examination carefully. Take notes in the space provided. If you need additional space for notes please use the backs of these pages or attach additional sheets. We are particularly interested in any significant differences between this examination and the way your agency conducts examinations in the field. We are also very interested in any insight you can provide about why the polygraph worked or did not work in this case. After you have finished watching the pretest interview, and before you watch the remainder of the examination, answer the questions following the notes section.

Pretest interview notes _____

Questions: Please indicate the response that most closely expresses your opinion by circling the number and writing the number in the blank at the left.

_____ In general, how much was this pretest like a pretest conducted by your agency in the field?

1 2 3 4 5 6 7
Not at all like the field Just like the field

_____ With regard to the length of the pretest, was this pretest longer, shorter, or the same length as a typical pretest conducted by your agency in the field?

1 2 3 4 5 6 7
Shorter About the Same Longer

Appendix B

---------- Was the examiner's description of the polygraph and the physiology of the detection of deception typical of those given by your agency in field examinations.

1 2 3 4 5 6 7
Less than the field About the Same More than the field

---------- Was the presentation and definition of the **RELEVANT** questions similar or dissimilar to the definition and presentation used by your agency in the field?

1 2 3 4 5 6 7
Not at All Like The Field Just like the field

If you felt that the definition and presentation of the **RELEVANT** questions was different from that used by your agency in the field please tell us about those differences.

---------- How well did the **RELEVANT** questions cover the subject's actions in the scenario?

1 2 3 4 5 6 7
Not Covered at All Covered Completely

---------- Was the emphasis placed on the **RELEVANT** questions in this pretest typical, less, or more than the emphasis placed on relevant questions during field examinations conducted by your agency?

1 2 3 4 5 6 7
Less emphasis About the Same More Emphasis

If you felt that the amount of emphasis placed on the **RELEVANT** questions was different from that used by your agency in field examinations, please tell us about the differences.

---------- Was the presentation and definition of the CONTROL questions similar or dissimilar to the definition and presentation used by your agency in the field?

1 2 3 4 5 6 7
Not at all like the field Just like the field

If you felt that the definition and presentation of the CONTROL questions was very different from that used by your agency in the field, please tell us about those differences.

---------- Did the control questions overlap the subject's actions in the scenario?

1 2 3 4 5 6 7
Complete Overlap No Overlap

---------- Was the emphasis placed on the CONTROL questions in this pretest typical, less, or more than the emphasis placed on relevant questions during field examinations conducted by your agency?

1 2 3 4 5 6 7
Less emphasis About the Same More Emphasis

If you felt that the amount of emphasis placed on the CONTROL questions was very different from that used by your agency in field examinations. please tell us about the differences.

Now please watch the intervals between the charts, and take notes if you wish.

Appendix B

Please answer the following questions by indicating the response that most closely expresses you opinion by circling the number and writing the number in the blank at the left.

---------- Were admonition about movement and/or breathing about the same as, stronger, or weaker than those given in the field?

1 2 3 4 5 6 7
Weaker Than the Field Same as the Field Stronger Than the Field

---------- In the between chart interval, were the **RELEVANT** questions discussed, and if they were, did the discussion emphasize the **RELEVANT** questions more, less, or about the same as they would be in a typical examination conducted by your agency?

1 2 3 4 5 6 7
Not discussed Less emphasis About the same More emphasis

---------- If you responded 'Not discussed' is that standard practice for your agency? YES NO

---------- In the between chart interval, were the **CONTROL** questions discussed, and if they were, did the discussion emphasize the **CONTROL** questions more, less, or about the same as they would be in a typical examination conducted by your agency?

1 2 3 4 5 6 7
Not discussed Less emphasis About the same More emphasis

---------- If you responded 'Not discussed' is that standard practice for your agency? YES NO

Given that the outcome of this examination was correct/incorrect, why do you think it turned out the way it did?

APPENDIX C
EXPERIMENT 1 QUESTIONNAIRE RESULTS

Affective Descriptors

Mean affective descriptors given by subjects on the Debriefing questionnaire (Appendix A) following their polygraph examinations are summarized in Table 17. Differences between the Innocent and Guilty/Knowledgeable subjects were tested with paired measurements t-tests, and the conditions were found to differ on the descriptors nervous, t (203) = 2.68, $p < 0.01$, tense, t (203) = 2.44, $p < 0.05$, guilty, t (203) = 7.12, $p < 0.001$, and anxious, t (203) =4.03, $p < 0.001$.

Table 17. Mean Affective Descriptors Given by Innocent and Guilty Subjects Following Their Polygraph Examinations.

Descriptor	Innocent	Guilty
Nervous	4.37	5.35*
Tense	4.06	4.98*
Guilty	2.03	4.66*
Anxious	4.02	5.64*
Fearful	3.36	3.70
Bored	3.48	3.57
Curious	8.13	8.57
Hopeful	6.86	6.66

*Indicates a significant difference between Innocent and Guilty conditions.

Data Reduction for the Attitude Questionnaire (Appendix A)

Questions were coded so that pro-polygraph responses were given scores of 1 or 2 and anti-polygraph responses were given responses of 4 or 5, neutral responses were coded as a 3. For example, answering 'always' (Choice a) to the question 'The polygraph is _____ able to tell when a person is lying' would score a 1, whereas choosing 'never' (Choice e) would score a 5.

Appendix C

Attitude Questionnaire (Appendix A)

Subjects' responses to the seven questions of the Attitude Questionnaire given before the subjects' examinations were subjected to a discriminant analysis to determine if the polygraph outcomes (Correct, Incorrect, Inconclusive) could be predicted from existing attitudes. The analysis failed to find a significant discriminant solution. That is, no response to any question, or responses to any combination of questions, predicted the outcomes of subsequent polygraph examinations.

The responses to the same questions during the post-test and follow-up administrations of the Attitude Questionnaire were analyzed with a Question (7) X Time (Post-Test, Follow-Up) X Outcome (Correct, Incorrect, Inconclusive) X Condition (Innocent, Guilty) RANOVA. That analysis revealed significant main effects for Outcome, F (2, 112) = 5.61, $p < 0.005$, and Question, F (6, 672) = 5.25, $p < 0.001$. There was a significant interaction of Outcome and Question, F (12,672) = 1.99, $p < .023$. The mean responses for question and outcome are shown in Table 18. .pa

Table 18. Mean responses to attitude questionnaire by test outcome collapsed across Post-Test and Follow-Up administrations.

Question Number	Outcome: Correct Mean	S.D.	(N)	Incorrect Mean	S.D.	(N)	Inconclusive Mean	S.D.	(N)
1.	2.0	.36	(71)	2.3	.59	(43)	2.2	.37	(8)
2.	1.9	.69	(71)	1.9	.66	(43)	1.9	.95	(8)
3.	2.0	1.0	(71)	2.4	1.3	(43)	2.1	.88	(8)
4.	1.6	.62	(71)	1.9	.85	(43)	1.8	.84	(8)
5.	1.9	.83	(71)	2.2	.79	(43)	2.1	.62	(8)
6.	2.6	.96	(71)	2.4	.83	(43)	2.2	.53	(8)
7.	2.2	.59	(69)	2.1	.73	(41)	2.0	.71	(8)

Possible changes in perceptions of test accuracy between the Pre-Test and the Post-Test administrations of the Attitude Questionnaire were tested with a Question (7) X Time (Pre-Test, Post-Test) X Outcome (Correct, Incorrect, Inconclusive) RANOVA. The hypothesis of primary interest was to examine whether a correct or incorrect outcome would interact with the subjects' perceptions of the polygraph accuracy. There was a significant Outcome by Question interaction, F (12,1164) = 1.82, $p < .04$, but there were no effects or interactions associated with the Time factor. The means of the seven questions collapsed across the Pre-Test and Post-Test Administrations of the Attitude questionnaire are presented by outcome are shown in Table 19.

Appendix C

Table 19. Mean responses to Attitude Questionnaire items by test outcome collapsed across Pre- and Post-Test administrations.

Question Number	Outcome: Correct Mean	S.D.	(N)	Incorrect Mean	S.D.	(N)	Inconclusive Mean	S.D.	(N)
1.	2.1	.39	(130)	2.2	.42	(58)	2.3	.37	(14)
2.	2.0	.75	(132)	1.9	.69	(59)	1.9	.79	(14)
3.	1.9	.89	(132)	2.1	.87	(59)	2.5	1.1	(14)
4.	1.5	.58	(132)	1.6	.61	(59)	1.9	.91	(14)
5.	1.9	.89	(132)	2.0	.75	(59)	2.2	.95	(14)
6.	2.4	.82	(130)	2.3	.65	(59)	2.1	.91	(14)
7.	2.0	.62	(129)	2.0	.52	(56)	2.0	1.0	(14)

Another issue tested in the Attitude Questionnaire data was whether or not certain tests (i.e., those containing control questions) were perceived as more intrusive than others. Question Number Six (How often does the polygraph violate a person's privacy?) at Time 2 was used as the dependent variable and Agency was a grouping variable for an ANOVA. This analysis found no difference between agencies in subjects' perceptions of how often the polygraph violates a person's privacy. The mean response to Question Six for the four agencies is shown in Table 20.

Table 20. Mean Post-Test response to Question Six by agency.

Agency	Mean	S.D.	N
MI	2.5	1.1	57
OSI	2.4	.77	51
CIA	2.5	1.1	44
NSA	2.3	.96	51

Percentage Questionnaire (Appendix A)

This questionnaire asked for percentage estimates of polygraph accuracy in various situations with different kinds of examines. The nine questions of the Percentage Questionnaire were subjected to discriminant analysis to determine if the

polygraph outcomes (correct, incorrect, inconclusive) could be predicted from existing attitudes. There was no significant discriminant solution. That is, no response to any question or any combination of responses to the questions, could predict test outcomes.

Subjects' percentage estimates of polygraph accuracy for various situations were analyzed in a Question (9) X Time (Pre-, Post-, and Follow-Up) X Outcome (Correct, Incorrect, Inconclusive) RANOVA. There was a significant Outcome effect, F (2, 110) = 4.64, $p < .012$, and a significant Outcome by Question interaction, F (16,880)=1.70,p=.041. The means for the nine questions collapsed across administrations are shown in Table 21 by outcome.

Table 21. Mean responses to Percentage Questionnaire by test outcome collapsed across administrations.

Question Number	Outcome: Correct Mean	S.D.	(N)	Incorrect Mean	S.D.	(N)	Inconclusive Mean	S.D.	(N)
1.	84.8	11.7	(92)	74.5	17.4	(21)	85.9	10.8	(7)
2.	84.9	12.7	(93)	76.6	17.8	(21)	88.1	11.6	(7)
3.	84.6	13.5	(93)	76.5	18.5	(21)	89.1	13.5	(7)
4.	82.0	15.9	(92)	72.2	17.0	(21)	76.0	16.9	(7)
5.	80.3	14.0	(92)	72.9	17.5	(21)	74.2	11.3	(7)
6.	83.7	12.7	(93)	75.3	18.5	(21)	83.2	10.8	(7)
7.	83.2	14.1	(93)	73.3	74.0	(21)	74.0	17.6	(7)
8.	87.1	12.0	(91)	75.1	20.1	(21)	87.0	12.0	(7)
9.	86.7	12.5	(90)	74.1	17.7	(21)	90.6	9.7	(6)

These results suggest that those who had incorrect outcomes rated the polygraph as less accurate than those with inconclusive or correct results. There was no effect for time or interaction between outcome and time.

Another theoretical question addressed with the Percentage Questionnaire data was the difference in perception of polygraph accuracy between others (Question One) and the subjects themselves (Question Nine). A paired t-test was used to compare the percentage estimates of polygraph accuracy 'in general' and 'on me (the subject)', and the difference was significant, t (201) = -4.37, $p < .001$. The means for the first and ninth questions were 81 and 84, the standard deviations were 16.6 and 17.1, respectively.