Polygraph

VOLUME 32 2003 NUMBER 4

Contents

Published Quarterly
© American Polygraph Association, 2003
P.O. Box 8037, Chattanooga, Tennessee 37414-0037

Polygraph

Subscription Information: *Polygraph* is published quarterly by the American Polygraph Association. Advertising and Editorial Address is P.O. Box 10342, Ft. Jackson, SC 29207 (USA). Subscription Rates: One year $80.00 (domestic), $100.00 (foreign). Change of address: APA National Office, P.O. Box 8037, Chattanooga, TN 37414-0037. THE PUBLICATION OF AN ARTICLE IN POLYGRAPH DOES NOT CONSTITUTE AN OFFICIAL ENDORSEMENT BY THE AMERICAN POLYGRAPH ASSOCIATION.

The Sheridan Press

Principles of Multiple-Issue Polygraph Screening A Model for Applicant, Post-Conviction Offender, and Counterintelligence Testing[1]

Donald J. Krapohl[2] and Brett A. Stern

Abstract

Multiple-issue polygraph testing is widely used in the US and other countries to uncover and monitor activities of individuals in circumstances where independent verification is difficult or impossible. This article discusses those factors that influence accuracy and utility of this singular application of polygraphy. Also discussed are decision rules for the polygraph, and how those decisions can be optimized, based on psychometric and statistical principles applied in large-scale medical screening.

Introduction

The use of the polygraph for routine multiple-issue testing of non-suspects is an extension of the polygraph's original role in criminal investigation. Multiple-issue polygraphy (MIP) can be traced back at least to the 1930s (Keeler, 1931), where it had some notable success in exposing widespread bank embezzlement that had not been detected by any other means. Even today, the main value of MIP is to uncover certain types of concealed past behaviors where there is very little information regarding the existence or prevalence of those behaviors among the examinee population. Despite longstanding concerns about the accuracy of polygraphy, and of MIP in particular, there are no other tools known by scientists to perform this function better (National Research Council, 2002).

The central topic of the present discussion is polygraph screening, one form of MIP, and it is important to explain the differences between the polygraph in the screening and diagnostic modes. Polygraph diagnostic examinations are those in which a specific incident known to investigators has taken place, and the polygraph is used to determine who may have participated in the incident. For example, if the event of interest were the loss of classified documents from a government office, investigators would polygraph those who could have been responsible for the loss, using test questions that focus on the missing documents. These tests, referred to here as single-issue tests (SITs), provide the most valid decisions possible in polygraphy.

In contrast to the SIT, a screening examination is one in which there is usually nothing known to the investigators that would prompt an investigation. Rather, a screening examination is used to determine whether members of a group had engaged in any of a number of proscribed activities. For example, the government has an interest in ferreting out bad security practices among employees who work in a highly classified environment. Those government employees could be routinely polygraphed on several of those practices,

[1] This article is one in a series under the heading Best Practices, and has simultaneous publication in *Polygraph*, the *Journal of the American Association of Police Polygraphists* and the *Journal of the Canadian Association of Police Polygraphists*. Copyright is retained by the American Polygraph Association.

[2] The opinions expressed in this article are those of the authors, and do not necessarily represent those of the US Government or the Department of Defense. Reprint requests should be sent to: Donald Krapohl, DoD Polygraph Institute, 7540 Pickens Ave., Ft. Jackson, SC 29207, or by e-mail to dkrapohl@aol.com.

such as illegal disclosures of classified information, mishandling classified documents, having covert dealings with foreigners, or deliberately damaging important equipment. Screening is also useful in other settings. Law enforcement agencies are entrusted by their citizenry to confirm that police applicants to whom they will grant special powers are not criminals, illegal drug users, or terrorists. Similarly, in the post-conviction sex offender domain the polygraph has been found exceptionally useful to therapists and supervisory officers to investigate whether the offender has relapsed into criminal sexual behavior or has engaged in precursor behaviors that precede these awful crimes. Often, the polygraph is the only means of uncovering hidden activities that threaten our families, our communities, and our country.

MIP can also be used for applications other than screening. Polygraphers sometimes use an MIP in criminal examinations to determine where to begin their single-issue testing. In a case of arson, the polygrapher could use the MIP to explore the possibility that the examinee started the fire himself, helped someone else start the fire, paid someone to start the fire, or whether he asked someone he knew to start the fire. This is sometimes called a multiple-facet examination, because it deals with a single event, but different aspects of the event. If the examinee shows significant reactions on specific questions, the subsequent questioning and testing would narrow to the areas covered by those questions. As with MIPs, however, the multiple-facet examination does not have the accuracy of the single-issue test (Blackwell, 1998) and is better suited to guiding subsequent testing than making a final determination in itself (Podlesny & Truslow, 1993).

The three applications of MIPs discussed earlier, applicant testing, sex offender monitoring, and counterintelligence screening, are the most commonly practiced in the US. These seemingly unrelated domains have elements in common that make MIP a highly valued tool. Among its greatest strengths is that MIPs encourage forthrightness. Faced with the prospect of having the truth revealed anyway, examinees appear to be more willing to volunteer information regarding socially proscribed acts they have committed in the past. It is commonly recognized that the greatest source of information regarding any person's past behavior is through self-report. Combining the polygraph with skilled questioning can promote candor in a way that simple interviewing cannot do.

Another reason for the MIP as a screening tool is that customers of polygraph screening are usually not interested in just a single type of behavior, but prefer to verify the presence or absence of several behaviors. MIPs overcome the impracticality of using single-issue polygraph testing for each of the 4-10 relevant areas typically covered in polygraph screening. MIPs are used as a first step to determine whether any of the relevant issues require further attention by the polygraph examiner, who can then do more interviewing and more testing on those issues. The MIP might also be used to guide the use of other investigative resources, enhancing the efficiency of that parallel process.

While screening polygraphy is typically thought of as a stand-alone method of deception testing, it is properly placed in context of a much larger process, where a failed screening test is used as a trigger for other investigative methods. This has much in common with medical screening, used to identify individuals in a community who harbor a very harmful disease. Medical screening is conducted in a very methodical way, using imperfect tools and methods in way that maximizes accuracy and utility.

There is benefit to the polygraph community in examining the manner in which the medical profession conducts screening. Using a recent example, suppose that doctors are very concerned about the spread of Severe Acute Respiratory Syndrome (SARS) by individuals passing through ports of entry into a country. SARS is a disease that is taken quite seriously, as it kills about 15% of those who contract it, and up to 55% of those in their 60s or older (Center for Infectious Disease Research & Policy, 2003). Unlike other diseases such as AIDS or hepatitis, which normally requires an exchange of bodily fluids to spread to another human, the risk factor for SARS is merely breathing air exhaled by someone already infected, making it far more contagious than most diseases. SARS has been blamed for more than 800 deaths, an untold number of those suffering secondary consequences, declines in stock market values, and a significant disruption in international travel (Drazen, 2003).

The base rate of SARS among trans-border travelers is miniscule, less than 1%. Given the virulence and potential harm of this disease to the general population, however, it is essential to identify every one of these rare individuals for isolation and treatment in order to prevent an epidemic. SARS can be diagnosed with high accuracy using a chest radiograph, a reverse transcription polymerase chain reaction (RT-PCR) test, and sputum Gram's stain, among other tests. While highly predictive of the presence of the SARS virus, these tests are expensive, intrusive, time intensive, or very inconvenient. It would not be practical to use these methods on any significant portion of the millions of travelers passing through a nation's airports. Medical experts instead developed a system that can identify those with a higher risk of carrying the disease using the "successive hurdles" approach (Meehl & Rosen, 1955). This method entails the use of not just one test, but a series of procedures, with each successive one having better specificity than the previous one.

A good example of that process is the one used in Hong Kong to detect SARS. It begins with a thermal scan of all entering and transiting passengers, looking for evidence of the fever that accompanies SARS-infected individuals. Those passengers having a body temperature lower than 38° C (100.4° F) are permitted to continue their travels without further testing. A higher temperature prompts an interview with the passenger regarding health- and travel-related issues, and is conducted by specially trained personnel. Those who provide answers to the questions, or have other indicators that point to an increased risk of SARS, are stopped from their travels and are transferred to a medical center for the tests that entail X-rays or collection of body fluids. Travelers who clear these tests are permitted to resume their flights. Those who continue to show evidence of SARS infection are given treatment in medical quarantine until they are found free of the disease.

The method used by the Hong Kong authorities minimizes disruption of travel for almost all travelers, while delivering the best diagnostic accuracy available. At each step, the concentration of possible SARS-infected travelers increases. As the level of intrusiveness increases, there are fewer travelers to be screened, and the most expensive and inconvenient tests are reserved for the very small group that shows the most elevated indicators of harboring the disease. Table 1 demonstrates how the low base-rate SARS-infected individuals might be culled from the larger population. The values here are arbitrary, and for illustration purposes only.

Table 1. *Illustration of the successive hurdles approach for SARS screening. Values are arbitrary.*

Prevalence a priori	Sample	Method	Positive results	True Positives	False Positives	Prevalence posteriori
.01%	100,000	Thermal Scan	500	10	490	2%
2%	500	Interview	200	10	190	5%
5%	200	Medical testing	11	10	1	91%

Let us now draw a parallel between medical and polygraph screening, using as an example an idealized model of law enforcement applicant screening. Police officers, because they are granted special powers under the law, must meet minimum standards of behavior. It is essential, for the sake of citizen safety, to prevent dangerous candidates from joining the ranks of sworn officers. Criminals such as murderers, terrorists, and child molesters probably have a very low base rate in the pools of police candidates, though given the enormous number of candidates who apply each year, the likelihood of these types of offenders entering the applicant pool is an actuarial certainty.

As part of the applicant processing, police candidates are frequently required to undergo polygraph screening and other tests for the purpose of uncovering potential problems with the candidate's suitability. The first phase of the polygraph examination is the multiple-issue polygraph test, where the test coverage will include questions about previous criminal conduct. Candidates who do well with the multiple-issue test are released from the polygraph session. Reactivity to a particular test question prompts more questioning by the polygrapher, who will try to solicit an explanation from the candidate regarding the adverse results. Sometimes the new information from the candidate is helpful to the hiring officials for making selection decisions, while in many other cases the cause of the reaction was something innocuous. Using the newly disclosed information the polygrapher will devise another polygraph test series using more focused questions. If the candidate passes this second testing, the session is over. However, if the candidate continues to do poorly, there may be other iterations of the test-interview-test cycle, with each succeeding test becoming increasingly narrow and questioning more direct and focused. The cycle ends when the candidate has passed the test, the organization has deemed that no more resources will be expended to polygraph the candidate, the candidate makes a disqualifying admission, or a final decision of deception is rendered by the testing examiner.

Unfortunately, polygraph screening in the field does not always follow this accepted model. Here are some of the larger shortcomings.

1. The most common problem is that some polygraphers never move beyond the multiple-issue screening phase. Decisions of DI are made on the multiple-issue test without bringing to bear more specific testing, and sometimes without affording examinees the opportunity to discuss their reactions on the screening test. This shortcut saves time and effort, but erodes the validity of the process.

2. Related to the problem of using a multiple-issue screening test as a stand-alone methodology is the use of trichotomous decisions: DI, NDI, and Inconclusive. In true screening, there are only two decisions: Negative (or NDI in polygraphy), and Tentative Positive (further processing required) Decisions of NDI, DI and Inconclusive are restricted only for diagnostic tests, meaning single-issue tests in polygraphy. Decisions from the screening phase are always dichotomous.

3. One of the principles of the successive hurdles approach that permits it to incrementally increase the accuracy of decisions is that a different method is used at each hurdle. Unfortunately, there are no post-polygraph technologies available to confirm positive screening results. The second best method is to alter the polygraph technique between the screening and diagnostic phases, so that the weaknesses of one method are not the same as those of the other method. Examiners who do follow-up testing, but use the same testing technique for all subsequent tests, have violated this principle.

4. Some polygraph screening methods have shown a low sensitivity to detecting deception. Decision rules must be used that do not permit

deceivers to avoid detection, even though this increases resource requirements to address truthtellers who are subjected to subsequent testing. If the initial test is not sufficiently sensitive to deception, the integrity and utility of the polygraph process will be compromised.

5. Some hiring officials, not understanding the limits of the polygraph, will press the polygraph examiner to ask an excess of test questions during the examination. It is well established in the literature that polygraph decision accuracy declines as a function of the number of different test questions. A polygraph exam that focuses on a single well-define area will afford the best accuracy, but have limited usefulness. A polygraph screening exam that covers dozens of relevant questions will perhaps garner more admissions, but it will also reduce accuracy to levels that make the polygraph little more than an interrogation prop. A compromise must be struck between accuracy and utility, but examiners should restrict topics to the smallest number possible.

While there are debates among polygraphers regarding the best screening test to use, it is important to note that there is no scientific evidence to suggest there are any differences among them in terms of accuracy. There are some critical factors beyond question sequences, though. Those factors are:

1. There are three essential types of questions: relevant, irrelevant, and comparison (if used). Other types of questions may be used for technical purposes, but evidence of their contribution to the accuracy of the technique is lacking. This does not preclude the use of symptomatic, sacrifice relevant, or similar questions, but in view of the current evidence, examinees and examinations do not appear to benefit from them.

2. Within certain limits, variation in question sequence is not crucial to the validity of the polygraph test. There are some very important principles to consider, however: the first two questions should not be relevant questions; no more than 3 relevant questions should be presented in a row; a relevant question should not be asked if the examinee is still reacting to the previous question; examinees must be advised in advance if there are to be repetitions of questions within a chart, and; cardio cuff pressure will restrict a chart to about 5 minutes or about 12 questions. There may be some small value in ensuring the first relevant question is also not the first evocative question, nor the first to be answered "no." This is because polygraph testing is actually an assessment of salience, and it naturally follows that valid testing requires that the salience of the relevant question should arise exclusively from the examinee's fear of detection, not from tertiary factors, and especially not those that are under the examiner's control. Use of the standard techniques is an excellent way of avoiding these types of problems.

3. Relevant test questions should be a few as possible. Accuracy is inversely related to the number of screening topics covered during a test. When a large number of screening topics is unavoidable, the questions should be broken into smaller groups of 2 - 5 questions, and given in separate tests.

4. Relevant questions should be as narrow in scope as possible. Restricting the scope of the relevant questions permits the examinee to be more confident in the accuracy of his answers, thereby enhancing the accuracy of the test.

5. Comparison questions, if used, should be as broad as possible. They should not explicitly cover the relevant area, though minor potential crossover should not invalidate the test. Replicated research has not yet found time bars to improve decision accuracy (Podlesny & Raskin, 1978; Horvath, 1988; Palmatier, 1991; Amsel, 1999.)

6. Relevant questions should focus on past behaviors. Avoid covering intent, attitudes, or future behavior.

7. There should be three or more presentations of each relevant question during testing.

8. All sessions should begin with an acquaintance test of some type (Kircher, Packard, Bell, & Bernhardt, 2001).

There are many existing multiple-issue screening tests that satisfy these principles: Air Force MGQT, Army MGQT, Navy MGQT, TES, RI, Utah screening test, or any of the various versions of the Exploratory ZCT, to name a few. There are also some lesser-known tests that may also be satisfactory.

No discussion of multiple-issue screening would be complete if it did not also consider the base-rate factor. It should be recognized that the base rate for deception is not equal in all settings. For example, the ratio of spies-to-applicant being polygraphed by the government is far leaner than the ratio of liars-to-truthtellers among post-conviction sex offenders undergoing sexual history polygraph examinations. These differences in base rates influence how much confidence one can have with a positive result arising from a screening test. When base rates of deception are very low, positive results are more likely to be erroneous than when base rates are high or balanced. To demonstrate this effect, Figures 1-3 show how often correct positive results occur in three base rate conditions. Decision accuracy and sample sizes are the same for all three conditions. Figure 1 starts with a balanced base rate, where half of the examinees are lying. In Figure 2, there are mostly liars being tested, such as in sexual history examinations of convicted sex offenders, and Figure 3 shows what happens when there are very few liars, such as in counterintelligence applicant screening.

Figure 1. Accuracy and errors with a balanced base rate, often seen in police applicant screening.

		Ground Truth	
		Truthful	Untruthful
Decision	Truthful	400	100
	Untruthful	100	400

Conditions

Accuracy	80%
Base Rate	50 out of 100
Number of Examinees	1000

Note: 20% of Untruthful decisions are false positives, 20% of Truthful calls are false negatives.

One of the sharpest criticisms against the screening polygraph is that it is biased against the truthteller. The evidence for this assertion is at best ambiguous, but it is so often repeated that to many it is taken as a fact. The argument is misinformative, however. Using the SARS-detection example again, it could be suggested from Table 1 that the thermal scan is biased against the SARS-free traveler because of a perceived false positive problem. At the low base-rate of SARS, false positives are overwhelming the

Figure 2. Accuracy and errors with a high base rate, such as in disclosure tests of convicted sex offenders.

		Ground Truth	
		Truthful	Deceptive
Decision	Truthful	40	190
	Untruthful	10	760

Conditions

Accuracy	80%
Base Rate	95 out of 100
Number of Examinees	1000

Note: Only about 1% of Untruthful calls are false positives, and 83% of Truthful calls are false negatives.

Figure 3. Accuracy and errors with a low base rate, such as routine counterintelligence screenings.

		Ground Truth	
		Truthful	Deceptive
Decision	Truthful	792	2
	Untruthful	198	8

Conditions

Accuracy	80%
Base Rate	1 out of 100
Number of Examinees	1000

Note: 96% of Untruthful calls are false positives, and less than 1% of Truthful calls are false negatives.

majority of the thermal scan results. If the debate stops at this point, as it often does among critics of the polygraph, one would come to the inescapable conclusion that the thermal scan should be abandoned because of its bias. However, this point of view ignores the context. A positive result for the thermal scanner brings about a focused interview, not immediate medical quarantine and treatment. Similarly, in the successive hurdles model a positive result for the multiple-issue polygraph test prompts a focused interview and more testing, not disqualification, termination, or imprisonment. Statements of accuracy or inaccuracy of the thermal scanner and the multiple-issue screening portion of the polygraph examination are meaningless when removed from the context of their respective roles in a larger process. This is one reason that research studies that look only at the multiple-issue screening test produce results that are likely to underestimate the accuracy of field practices when a successive hurdles approach is used.

In a similar vein, the assertion that the imperfect accuracy of the polygraph applicant screening process hurts the chances of truthtellers is based on an incomplete understanding of the conditions under which the polygraph is normally used. Despite assertions to the contrary, polygraph screening actually improves the hiring prospects for qualified candidates. The following thought experiment explains why this is so.

Suppose a large police department had openings for 100 new officers, and had no polygraph screening program. Also assume that there were 200 applicants, with 25% of them having engaged in past behavior that would have made them unqualified by department hiring standards, such as drug use, theft, violence, sexual abuse, or fraud. If the police department randomly chose from those candidates, it would place 75 qualified and 25 unqualified applicants into those 100 positions.

Let us now look at another police department with similar conditions and requirements, but one that had a polygraph screening program that produced a very modest accuracy of, say, 70% against a chance accuracy of 50%. If this second police department used the polygraph to help select out unqualified candidates, and only chose from among those who passed the polygraph, it would place 87 qualified and 13 unqualified into the 100 positions. These numbers relate only to hiring decisions that are based on polygraph results, and are irrespective of any disqualifying admissions a candidate may make during the polygraph session. The use of the polygraph, even with its flawed accuracy, would increase the opportunities for qualified candidates by 12 positions per hundred, and reduce those of unqualified candidates by an equal number. Rather than discriminating against truthful candidates as has been alleged by some, a well designed polygraph program would significantly improve the chances for qualified police applicants. Incremental increases in polygraph decision accuracy, such as what can be achieved with the successive hurdles approach and a thoughtful use of the polygraph to encourage candor, may further boost the proportion of qualified applicants selected. A thorough discussion of this largely ignored effect, and how it helps qualified candidates, can be found in an article by Martin and Terris (1991).

One final consideration is the cost of errors. These costs differ by application. A false positive error for an applicant could mean that he needs to submit his resume to other organizations. While this may be inconvenient, it does not entail the loss of something for which he had automatic entitlement. In contrast, a false positive decision in the post-conviction sex offender arena can, in some jurisdictions, contribute to the probationer being unjustly remanded back to prison, or unfairly prevent reunification with unvictimized family members. It would be very difficult to reconcile this consequence with professional ethics if it were the result of a polygraph examiner's decision based on a multiple-issue examination only. In counterintelligence screening of current employees, the US government has long recognized the value of the successive hurdles approach, and there are no adverse personnel actions taken that are based solely on a failed personnel screening examination. In that environment, even those who fail the diagnostic phase of the polygraph process on behaviors with extremely low base rates, absent other corroborating evidence these employees would suffer no sanctions.

Summary

Following the medical screening model, it is possible to increase the efficacy of multiple-issue polygraph screening using the successive hurdles approach. For this method to be effective:

1. The initial screening test must be sensitive to deception, even though there will be some truthful examinees who will find themselves being given additional testing along with the deceptive examinees. If the initial screening test is insensitive to deception, deceptive examinees will slip through the system with little likelihood that these errors will be caught before the employee has committed some act that brings them to the attention of managers.

2. When an examinee is found truthful to the screening phase, the polygraph decision will be NDI, or its equivalent. For examinees not clearing the screening phase, it is not appropriate to make a DI call, or its equivalent. When an examinee has not passed the screening phase (MIP), more discussion and testing are warranted, followed by testing with questions of narrower scope. At the end of the diagnostic testing phase, calls of NDI, DI, or Inconclusive can be made. Decisions of countermeasures (or non-cooperation) may also be made when procedures are in place to detect them, and policies are developed to address those cases.

3. When it is necessary to do follow up testing after a screening test, it is best to use a different technique. Changing techniques can minimize the carryover of the same errors from screening to diagnostic phases, and also make countermeasures more difficult for the examinee to conceal.

Acknowledgements

We are grateful to Dr. John Kircher of the University of Utah who originally suggested to the authors the benefits of the successive hurdles model. We also appreciate the helpful suggestions offered by Raymond Nelson to an earlier draft of this article.

References

Amsel, T.T. (1999). Exclusive or Nonexclusive comparison questions: A comparative field study. *Polygraph, 28*(4), 273-283.

Blackwell, N.J. (1998). PolyScore 3.3 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations. Report DoDPI97-R-006. DTIC AD Number A355504/PAA. Department of Defense Polygraph Institute: Ft. McClellan, AL.

Center for Infectious Disease Research & Policy (2003). Estimates of SARS death rates revised upwards. Available at http://www.cidrap.umn.edu/cidrap/content/hot/SARS/news/may0703SARS.html

Drazen J.M. (2003, July 24). SARS — Looking back over the first 100 days. *The New England Journal of Medicine, 349*(4), 319-320.

Horvath, F. (1988). The utility of control questions and the effects of two control question types in field polygraph techniques. *Journal of Police Science and Administration, 16*(3), 198-209.

Kircher, J.C., Packard, T., Bell, B.G., & Bernhardt, P.C. (2001). Effects of prior demonstrations of polygraph accuracy on outcomes of probable-lie and directed-lie polygraph tests. DoDPI02-R-0002. DTIC AD Number A404128. University of Utah.

Martin, S.L., & Terris, W. (1991). Predicting infrequent behavior: Clarifying the impact on false-positive rates. *Journal of Applied Psychology, 76*(3), 484-487.

Meehl, P.E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, and cutting scores. *Psychological Bulletin, 52*(3), 194-216.

Palmatier, J.J. (1991). *Analysis of Two Variations of Control Question Polygraph Testing Utilizing Exclusive and Nonexclusive Controls.* Unpublished masters degree thesis, Michigan State University, East Lansing, Michigan.

Podlesny, J.A., & Raskin, D.C. (1978). Effectiveness of techniques and physiological measures in detection of deception. *Psychophysiology, 15*(4), 344-359

Podlesny, J.A., Truslow, C.M. (1993). Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology, 78*(5), 788-797.

Suggested Readings

Crewson, P.E. (2001). A comparative analysis of polygraph with other screening and diagnostic tools. Report to the Department of Defense Polygraph Institute, contract no. DABT60-01-R-0003. Electronic copies available from the first author (dkrapohl@aol.com).

Swets, J.A., Dawes, R.M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1-26.

CRS Report for Congress

Polygraph Use by the Department of Energy: Issues for Congress[1]

Alfred Cumming[2]

Summary

In the aftermath of the Wen Ho Lee case and the growing concern over the Department of Energy's (DOE) counterintelligence program that followed, DOE in March 1999 began developing its first-ever polygraph screening program affecting an estimated 800 DOE employees with access to sensitive and classified information.

Congress in October 1999 mandated DOE polygraph testing (P.L. 106-65, Sec. 3154) and expanded the program to cover 13,000 DOE employees with access to sensitive and classified information. The following year, Congress further expanded polygraph testing to cover approximately 20,000 DOE employees (P.L. 106-398, Sec. 3135) with the addition of new eligibility categories. In part because of continuing opposition by some DOE nuclear weapons laboratory employees, Congress in 2001 requested that the National Academy of Sciences (NAS) review the scientific evidence regarding the validity and reliability of the polygraph, particularly when used for personnel security screening. Congress directed DOE to institute a new polygraph program based upon the NAS findings (P.L. 107-107, Sec. 3152).

NAS completed its study in October 2002, concluding that while polygraph testing is more effective when used in connection with event-specific investigations, its accuracy is insufficient to justify reliance on its use in screening current and prospective federal agency employees — DOE's principal purpose in using the polygraph. According to NAS, in populations such as DOE's, where there is an extremely low level of major security violations, the polygraph has serious limitations for use in security screening to identify security risks. NAS also reported that there is a realistic possibility that the polygraph might be defeated with countermeasures.

Although acknowledging the NAS findings, Energy Secretary Spencer Abraham announced on April 14, 2003, that DOE would continue to use the polygraph for screening purposes, citing it as an effective component of DOE's counterintelligence program. He said that DOE does not use the polygraph on a stand-alone basis but as part of a larger fabric of investigative and analytical reviews to help security personnel determine if someone is suitable to access to classified data. He also asserted that polygraphs have value in deterring unauthorized disclosures of classified information.

Pointing to the NAS findings, some Members of Congress have called on the Energy Secretary to review his decision, and have expressed a desire for a more focused polygraph program that would subject fewer DOE employees to testing. Others have cautioned that a false sense of confidence can arise from reliance on a technique they believe is inaccurate. They also cited NAS's warning that the polygraph can be defeated by countermeasures.

There are several possible approaches Congress might assess, including retention of the status quo, the establishment of a more focused polygraph program under which those occupying only the most sensitive positions would be polygraphed; more research into alternatives to the polygraph; and the elimination of the polygraph for screening purposes altogether. This report will be updated as warranted.

Introduction

Since its establishment in 1977, the Department of Energy (DOE) has been frequently criticized for adopting a lax approach to countertintelligence (CI), particularly in connection with its nuclear weapons laboratories.[3] Years of increasingly critical CI reviews culminated in 1998 when intelligence evidence suggested that the People's Republic of China (PRC) had stolen secrets from DOE's national security laboratories.[4] President Clinton responded by issuing Presidential Decision Directive (PDD) 61, which fundamentally restructured DOE's CI program. The President directed DOE to develop and implement specific security measures, including the possible use of polygraph testing, to reduce the threat to classified information.

In March 1999, DOE began to develop a CI-scope polygraph to screen employees occupying security-sensitive positions. Questions asked as part of a CI-scope polygraph are limited to topics concerning the individuals' involvement in espionage, sabotage, terrorism, unauthorized disclosure of classified information, unauthorized foreign contacts, and deliberate damage to or malicious misuse of a U.S. Government information or defense system. In August 1999, DOE proposed expanding the CI polygraph program to contractor and Federal employees at its facilities in positions with access to the most sensitive categories of classified information and materials, as well as to applicants for such positions.[5] In October 1999, Congress mandated what until then had been a DOE-discretionary polygraph testing program (P.L. 106-65, Section 3154). Congress also expanded the number of those required to take the polygraph to 13,000.[6] In December 1999, Energy Secretary Bill Richardson announced that CI interests could be satisfied with approximately 800 polygraphs.[7] DOE originally had intended its program to cover approximately 3,000 employees, but the number was reduced after protests from national laboratory employees.[8] Richardson said he would seek legislation to ensure consistency between DOE's implementation plan and congressional direction.[9] Instead, Congress in 2000, prompted by continuing security concerns, approved legislation (P.L. 106-398, Section 3135) further expanding, by statute, the program to cover approximately 20,000 DOE employees.[10] By subsequently eliminating overlapping categories of covered employees, DOE today polygraphs approximately 16,000 employees.

In 2002, Congress instructed the National Academy of Sciences (NAS) to review the scientific evidence regarding the polygraph's validity and reliability and directed DOE to institute a new program based upon NAS findings (P.L. 107-107, Section 3152). Congress said a new program should "minimize the potential for release or disclosure of classified data, materials, or information."

Background

Debate continues over the validity and reliability of the modern polygraph, first developed early in the 1900s. What is undisputed is that the polygraph machine does not measure deception but rather is an instrument that charts changes in a individual's respiration, heart rate, blood pressure, and sweat gland activity in response to a series of yes/no answers.[11] Polygraph examiners determine whether a person's physiological reaction is stronger in responding to certain questions when contrasted with recorded reactions to a series of comparison "control" questions. Stronger reactions indicate that the individual may be deceptive. It is these physiological responses which are at the heart of the ongoing debate over the validity of polygraph testing.[12] Scientists studying the polygraph further note a distinction between the polygraph test and the polygraph examination, which includes the test and the interrogation surrounding it. The first represents an attempt to capture accurate psychophysiological indicators of deception. The second is a tool for revealing truth.[13]

The polygraph is used for three principal purposes: event specific or exculpatory — for example, when a crime has been committed; preemployment screening; and screening current employees. The Intelligence Community uses the polygraph as a screening device and investigative tool. The

Department of Defense (DOD) uses it almost exclusively as an investigative tool. DOD does use polygraphs for employee screening, but only for individuals granted exceptional clearances for highly sensitive programs.[14]

Although DOE has long used the polygraph as an investigative tool, only since 1999 has it employed it as a screening tool. The Energy Department turned to polygraph testing after President Clinton issued PDD 61 in response to long-standing concerns about security at DOE weapons labs and specifically because of intelligence evidence indicating that the PRC may have stolen secrets from DOE's weapons labs. The President directed DOE to consider establishing a polygraph program as one component of a comprehensive CI program that could include background checks, periodic reinvestigations, monitoring of financial records, restrictions on publishing materials, and, for some employees, mandatory drug testing and medical assessments.[15] Under current DOE regulations, neither DOE nor its contractors may take an adverse personnel action against an individual solely on the basis of a polygraph result indicating deception.[16]

DOE cited three principal reasons when it proposed polygraph screening in 1999.[17] First, a CI-scope polygraph, according to DOE, serves both as means to deter unauthorized disclosures of classified information and to detect early any disclosure of classified or sensitive information. The latter, DOE argues, allows it to promptly take steps to mitigate any damage to the national security. Second, DOE suggested that the polygraph examination is essential in granting interim personnel security clearances on an expedited basis. Finally, DOE argued that a polygraph examination provides an important tool that is available upon employee request to expeditiously resolve any outstanding issues in a CI or personnel security investigation.

Some See Polygraph's Utility But Many DOE Scientists Are Skeptical

Many DOE laboratory personnel have a "very negative" attitude towards the polygraph, according to the report of the Redmond Panel, a panel of experts which reviewed DOE CI capabilities at DOE's national security laboratories on behalf of the House Permanent Select Committee on Intelligence.[18] The attitude toward polygraphs at the laboratories, according to panel findings, runs the gamut from cautiously and rationally negative to emotionally and irrationally negative.[19] The Panel noted in its findings that never before have so many cleared employees of a government organization had to have their clearances threatened by the institution of the polygraph.[20] The Panel also noted that scientists do, in fact, represent a particular problem with regard to the administration of polygraphs. "They are most comfortable when dealing with techniques that are scientifically precise and reliable," according to the Panel. "The polygraph, useful as it is as one of several tools in a CI regimen, does not meet this standard. Accordingly, many scientists who have had no experience with it are skeptical of its utility."[21] The Panel went on to note, however, that "...polygraphs, while not definitive in their results, are of significant utility in a broader comprehensive CI program. The polygraph is an essential element of the CI program and it will not work until it is accepted by those who are subject to it."[22] In its report, NAS said that polygraph testing has some utility in "deterring security violations, increasing the frequency of admissions of such violations, deterring employment applications from potentially poor security risks, and increasing public confidence in national security organizations....Such utility derives from beliefs about the procedure's validity, which are distinct from actual validity or accuracy."[23]

The Society of Professional Scientists and Engineers, an association of current and retired scientists at Lawrence Livermore National Laboratory, argues that the polygraph is not only scientifically invalid and unreliable but lacks utility as well. "Their unreliability renders polygraphs incapable of catching spies and can lead to false accusations of innocent workers who may find themselves defenseless against the machine's oscillations," according to the Society.[24] Other critics argue that the polygraph has failed to uncover such prominent spies as Aldrich Ames and indeed maintain that spies know how to employ countermeasures against the polygraph.

Dearth of Scientific Evidence Underlying the Polygraph

As distinct from the utility of the polygraph, supporters and critics of the polygraph agree that the scientific evidence relevant to the accuracy of polygraph screening tests — the principal purpose of DOE's polygraph program — is extremely limited. The NAS said it found only one flawed field study that provided evidence directly relevant to accuracy for preemployment screening.[25] The American Polygraph Association (APA), the largest polygraph association consisting of examiners in the private, law enforcement, and government fields, blames the paucity of research into the scientific basis for the polygraph or any other deception detection technique on a lack of resources.[26] The NAS agreed, noting that the lack of serious investment in such research is "striking," given the heavy reliance of the government on the polygraph, especially for screening for espionage and sabotage. [27]

What The Available Evidence Does Show

The NAS, in its study, arrived at a number of conclusions. First, it concluded that polygraph testing, particularly with regard to personnel screening, yields an unacceptable choice for DOE employee security screening between too many loyal employees falsely judged deceptive and too many major security threats left undetected. The polygraph's accuracy, according to the NAS, in distinguishing actual or potential security violators from innocent test takers is insufficient to justify reliance on its use in employee security screening in federal agencies.[28]

Second, the NAS concluded that, based upon field reports and indirect scientific evidence, polygraph screening may have some utility for achieving such objectives as deterring security violations, increasing the frequency of admissions of such violations, deterring employment applications from potentially poor security risks, and increasing public confidence in national security organizations. Such utility, however, derives from beliefs about the validity of the procedure, and are distinct from "actual validity or accuracy," according to NAS.[29] The NAS also concluded that the proportion of spies, terrorists and other major national security threats among the employees subject to polygraph testing in the DOE labs presumably is very low, and that polygraphs therefore should not be counted on for detection when screening population with low rates of the target transgressions, because "screening populations with very low rates of the target transgressions (e.g., less than 1 in 1,000) requires diagnostics of extremely high accuracy, well beyond what can be expected from polygraph testing."[30] NAS also stated that countermeasures pose a potentially serious threat to the performance of polygraph testing because all the physiological indicators measured by the polygraph can be altered by conscious efforts through cognitive or physical means. NAS noted that "there is enough empirical evidence to justify concern that successful countermeasures may be learnable."[31]

The NAS findings essentially track the results of a similar research review completed by the Congressional Office of Technical Assessment (OTA) in 1983, which concluded that there was not adequate evidence at that time to establish the scientific validity of the polygraph test for personnel security screening. OTA went on to more broadly state that the establishing the overall validity of the polygraph is not possible. OTA cited two reasons. First, the polygraph is a very complex process that is much more than the instrument. The types of individuals tested, examiner's training, purpose of the test, and types of questions asked, among other factors can differ substantially, according to the OTA report. Second, OTA noted, the research on polygraph validity varies widely in terms of results and the quality of the research design and methodology. "... conclusions about scientific validity can be made only the context of specific applications and even then must be tempered by the limitations of available research evidence," according to OTA.[32]

Polygraph supporters such as the APA in turn cite 80 research projects, published since 1980, showing accuracy ranges for the polygraph from 80 to 98 percent.[33] While conceding that there has been only a limited number of research projects on the accuracy of polygraph testing for screening, the APA argues that "real world conditions are difficult if not impossible to replicate in a mock crime or laboratory environment for the purpose of assessing effectiveness."[34] The APA further argues that the same physiological measures are recorded and the same basic psychological principles may apply in both the event specific and pre-employment screening examinations and therefore there is no reason to believe that there is a substantial decrease in the accuracy rate for the preemployment circumstance. The few studies that have been conducted on pre-employment testing support this contention, according to the APA.[35]

U.S. intelligence community agencies, however, continue to believe the polygraph is a useful screening tool. The CIA claimed to have classified research to support their use of polygraph tests but would not share it with OTA at the time of its study. According to OTA, in its 1983 report, CIA and NSA use the polygraph not to determine deception or truthfulness per se, but as a technique of interrogation to encourage admissions. OTA reported that the polygraph examination results that are most important to NSA security adjudicators are the data provided by the individual during the pre-test or post-test phase of examination. The Director of Central Intelligence Security Committee concluded that polygraph was the most productive of all background investigation techniques. But OTA pointed out that the study was one of utility, not validity.

DOE Proposes To Maintain Current Polygraph Program

The National Defense Authorization Act for Fiscal Year 2002 (P.L. 107-107, Section 3152) directed DOE to issue a notice of proposed rule-making for a new polygraph program based upon the findings of the NAS polygraph review. The Act also stated that the purpose of any such new program would be to minimize the potential for release or disclosure of classified data, materials, or information.

On April 14, 2003, DOE, to satisfy the congressional directive, published a notice of proposed rule-making "to begin a proceeding to consider whether to retain or modify [DOE's] current Polygraph Examination Regulations."[36] Secretary Abraham made clear that DOE intended to retain polygraph screening as a component of the Department's CI program. In doing so, he acknowledged NAS's recommendation against using the polygraph for employee screening and the congressional directive that he take NAS's views into account. But he said that maintaining polygraph testing was "consistent with the statutory purpose of minimizing the risk of disclosure of classified data."[37] He also said that DOE's use of the polygraph only as a trigger for a detailed follow-up investigation, and not as a basis for personnel action, was compatible with NAS's conclusion that if polygraph screening is to be used at all, it should be used in this fashion.[38]

Critics of the Secretary's decision, including Senator Bingaman (D-NM), said relying on a technique as inaccurate as the polygraph could produce a false sense of confidence. That overconfidence, Bingaman suggested, "can be the real danger to national security." Applying polygraphs to employee screening could lead to either too many loyal employees who will judged deceptive, or too many major security threats undetected, Bingaman noted.[39] Senator Pete Domenici said, "I continue to believe that the system is too much an affront[,] especially since the polygraph program was so thoroughly criticized by the National Academy of Sciences. I hope the department will rethink this situation."[40]

Issues for Congress

A More Focused Polygraph

One issue for Congress is whether DOE's polygraph screening program should focus on a smaller number of individuals occupying only the most sensitive positions. Approximately 16,000 DOE employees, falling into eight separate categories, currently are polygraphed. Critics argue that the program requires a screening polygraph for virtually every DOE employee and contractor who holds a security clearance, without regard to the level of sensitivity of that persons's activities or access. Such a program exposes a large population to polygraph examination without regard to the risk associated with that person's access. One result, according to critics, is that polygraphs have caused, and may continue to cause, severe morale problems and thereby ultimately undermine the very goal of good security.[41]

DOE counters by saying that the Department's classified information consists in significant measure of information regarding nuclear weapons of mass destruction and that the consequences of compromise could be profoundly significant. DOE contends it is under a particular obligation to make sure that no action that it takes be susceptible to misinterpretation as a relaxation of controls over information concerning these kinds of weapons.[42]

Additional Research

Critics and supporters alike agree that further research into the scientific basis for psychophysiological detection of deception by any technique is warranted. The NAS, in its report, suggests that if the government continues to rely heavily on the polygraph, some research effort should focus on putting the polygraph on a firmer scientific foundation. NAS cautions, however, that given the inherent ambiguity of the physiological measures used in the polygraph suggest that further investments in improving polygraph technique and interpretation will bring only modest improvements in accuracy.[43] Rather, NAS recommends the development of a broader research program to detect and deter security threats be developed, rather than focus on polygraph research.[44] NAS points out that potential alternative techniques such as measurements from brain activity and other physiological indicators, facial expressions, voice quality and other aspects of demeanor show some promise, but that "none has yet been shown to outperform the polygraph. None shows any promise of supplanting the polygraph for screening purposes in the near term."[45] According to NAS, any such research program should be largely administered by "an organization or organizations with no operational responsibility for detecting deception and no institutional commitment to using or training practitioners of a particular technique."[46]

NAS pointed out two particular areas worthy of more research — computerized analysis of polygraph records to improve the accuracy of test results by using more information from polygraph records than is used in traditional scoring methods; and combining polygraph information with information from other screening techniques,. NAS also concluded that more research needs to be conducted with regard to countermeasures, but pointed out that policy makers must weigh the danger of public knowledge of countermeasures against the benefits of a robust public research program.[47]

Supporters, while claiming that the polygraph now provides satisfactory detection and deterrence, favor enhanced research efforts on grounds that they will certainly expand the capacity to improve the polygraph's validity and reliability.[48] At the same time, supporters note that real world conditions are difficult if not impossible to replicate in a mock crime or laboratory environment for the purpose of assessing the polygraph's effectiveness. A lack of resources, according to supporters, also has impeded research efforts.

Responding to the NAS research recommendation, the Senate Select Committee on Intelligence in its fiscal year 2004 intelligence authorization bill authorizes $500,000 for the National Science Foundation and the Office of Science and Technology to lead a more focused research effort on alternatives to the polygraph.[49]

Discard Use Of Polygraph For Screening

Another issue for Congress is whether to discard the use of polygraph screening — as opposed to event specific use where the accuracy is well above chance but below perfection [50] — altogether. Critics label the screening polygraph as misguided and suggest that it be shelved in favor of more thorough examination of staff's financial records and travel, and more frequent reinvestigation by traditional means. These critics argue that the screening polygraph gives a false and dangerous sense of over confidence to authorities that they have adequately screened for spies.[51] They believe that this, in turn, could lead to a relaxation of other methods of ensuring security, such as periodic security re-investigation and vigilance about potential security violations in facilities that use the polygraph for employee security screening.[52] Critics also argue that polygraph test accuracy can be undermined by countermeasures, seriously undermining the any value of polygraph security screening.[53]

Supporters argue that the polygraph is the best tool currently available to detect deception and that it remains an important tool to detect deception in selected national security and law enforcement matters. Some supporters distinguish between the polygraph's utility and its scientific validity and reliability. While not definitive in its results, they argue, the polygraph is of significant utility in a broader comprehensive CI program.[54] Some government organizations further claim to have classified information which supports their use of polygraph tests.[55] And virtually all supporters of polygraph screening argue that polygraph testing is just one tool among several used as part of comprehensive CI program.

Endnotes

[1] Downloaded through the Congressional Research Service Web. Originally published on July 8, 2003.

[2] Specialist in Intelligence and National Security; Foreign Affairs, Defense and Trade Division

[3] DOE has three nuclear weapons laboratories where classified nuclear weapons research is conducted: Los Alamos National Laboratory, Los Alamos, NM; Lawrence Livermore National Laboratory, Livermore, CA; and Sandia National Laboratories, Albuquerque, NM and Livermore, CA.

[4] For a comprehensive review of this issue, see CRS Report RL30143, *China, Suspected Acquisition of U.S. Nuclear Weapons Secrets*, by Shirley Kan. See also *Attorney General's Review Team on the Handling of the Los Alamos Laboratory Investigation*, May, 2000, at [http:\\www.FAS.org].

[5] *Federal Register* 64, no. 242 (Dec. 17, 1999), p. 70963.

[6] See United States Department of Energy News, *DOE Polygraph Implementation Plan Announced*, Dec. 13, 1999.

[7] Ibid

[8] Andrea Widner, "DOE Lab Employees Protest New Law Mandating Polygraph Tests," *Knight Ridder/Tribune News*, Nov. 9, 2000. 7

[9] See United States Department of Energy News, *DOE Polygraph Implementation Plan Announced*, Dec. 13, 1999.

[10] *Federal Register* 68, no. 71 (April 14, 2003), p. 17887. [2]

[11] A polygraph instrument will collect physiological data from at least three systems in the human body. Convoluted rubber tubes that are placed over the examinee's chest and abdominal areas will record respiratory activity. Two small metal plates, attached to the fingers, will record sweat gland activity, and a blood pressure cuff, ro similar device will record cardiovascular activity.

[12] See the National Research Council of the National Academy of Sciences, *The Polygraph and Lie Detection*, 2002, p. 13

[13] Ibid., p. 21.

[14] See Commission on Science and Security, *Science and Security in the 21st Century, A Report to the Secretary of Energy on the Department of Energy Laboratories*, Apr., 2002, p. 54.

[15] *Federal Register* 64, no. 242 (Dec. 17, 1999), p. 70962.

[16] Ibid., p. 70962.

[17] Ibid., p. 70963.

[18] U.S. Congress, House Permanent Select Committee on Intelligence, *Report of the Redmond Panel*, June 21, 2000, pp. 7-8.

[19] Ibid., p. 7.

[20] Ibid., p. 7.

[21] Ibid., p. 8.

[22] Ibid., p. 8.

[23] See the National Research Council of the National Academy of Sciences, *The Polygraph and Lie Detection*, 2002, p. 6

[24] Society of Professional Scientists and Engineers, *SPSE Speaks Out on Polygraphs*, Aug. 13, 1999.

[25] See the National Research Council of the National Academy of Sciences, *The Polygraph and Lie Detection*, 2002, p. 3.

[26] American Polygraph Association, *Statement of the American Polygraph Association Pertaining to the National Academy of Sciences (NAS) Report on the Use of the Polygraph*, undated

[27] See the National Research Council of the National Academy of Sciences, *The Polygraph and Lie Detection*, 2002, p. 8.

[28] Ibid., p. 6

[29] Ibid., p. 8.

[30] Ibid., pp. 5-6.

[31] Ibid., p. 216.

[32] Office of Technology Assessment, *Scientific Validity of Polygraph Testing,* Nov. 1983, p. 4.

[33] American Polygraph Association, *Polygraph Issues and Answers*, undated.

[34] American Polygraph Association, *Statement of the American Polygraph Association Pertaining to the National Academy of Sciences (NAS) Report on the Use of the Polygraph*, undated.

[35] American Polygraph Association, *Polygraph Issues and Answers*, undated.

[36] *Federal Register* 68, no. 71, p. 17886.

[37] U.S. Department of Energy, "DOE Issues Notice of Proposed Rulemaking on Polygraph Use," press release, Apr. 14, 2003.

[38] Ibid.

[39] See press statement of Senator Bingaman, Apr. 14, 2003

[40] See news release of Senator Domenici, "Domenici: DOE Worries Shouldn't Mean Continuation of Flawed Polygraph Policy," Apr. 15, 2003.

[41] See Commission on Science and Security, *Science and Security in the 21st Century, A Report to the Secretary of Energy on the Department of Energy Laboratories*, April 2002, pp. 55-56. The Commission points out in its report that DOD does use the polygraph for screening purposes, but only for individuals granted exceptional clearances for highly sensitive programs. In the case of the Intelligence Community, according to the Commission, the polygraph is an integral — "and more important, an accepted — part of the intelligence community's security practices and culture. People are aware of this practice before accepting employment in intelligence organizations, and they accept it as an integral part of a more comprehensive security architecture."

[42] *Federal Register* 68, no. 71, pp. 17888-17889.

[43] See the National Research Council of the National Academy of Sciences, The Polygraph and Lie Detection, 2002, p. 213.

[44] Ibid., p. 9.

[45] See the National Research Council of the National Academy of Sciences, *The Polygraph and Lie Detection*, 2002, p. 8.

[46] Ibid., p. 229.

[47] Ibid., p. 231.

[48] See American Polygraph Association, *Statement of the American Polygraph Association Pertaining to the National Academy of Sciences (NAS) Report on the Use of the Polygraph,* undated.

[49] See S.Rept. 108-44, p. 29.

[50] See the National Research Council of the National Academy of Sciences, The Polygraph and Lie Detection, 2002, p. 4.

[51] See comments by the Society of Professional Scientists and Engineers to proposed polygraph examination regulations, 10 CFR Pat, 709, Federal Register 68, p. 17886, Apr. 14, 2003.

[52] See the National Research Council of the National Academy of Sciences, The Polygraph and Lie Detection, 2002, p. 7

[53] Ibid., p. 5.

[54] U.S. Congress, House Permanent Select Committee on Intelligence, *Report of the Redmond Panel*, June 21, 2000, pp. 7-8.

[55] See Office of Technology Assessment, *Scientific Validity of Polygraph Testing,* Nov. 1983, p. 100. OTA said it did not review this research.

Comparison of Skin Conductance and Skin Resistance Measures for the Detection of Deception

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

Abstract

Traditional analog polygraph instruments typically record skin resistance (SR), whereas academic psychophysiologists typically record skin conductance (SC) and have argued that SC is superior to SR. The present study tested if SC or SR is more useful for the detection of deception. 336 participants in a previous experiment (DODPI97-P-0016) were tested about their participation in a mock theft. Half of the participants were guilty of committing the theft, and the other half were innocent. Probable-lie polygraph tests (PL) were administered to half of the innocent and half of the guilty participants, and directed-lie tests (DL) were administered to the remaining participants. Participants were paid $30 and were offered an additional $50 to convince the polygraph examiner of their innocence. Digitized recordings of SC subsequently were transformed to SR. A computer measured the amplitudes and other features of SC responses and SR responses to comparison and relevant test questions.

Both SC and SR were highly diagnostic of truthfulness and deception, and no evidence was found to favor either SC or SR for either PL or DL polygraph tests. Univariate analyses revealed no differences between SC and SR in their ability to discriminate between truthful and deceptive individuals, and multivariate analysis indicated that either measure might be used in combination with other physiological measures to detect deception. However, these conclusions apply only to SR recordings obtained when a constant-voltage is applied to wet electrodes. Additional research would be needed to compare such laboratory-grade measurements of SC and SR to electrodermal activity as it is measured by traditional analog and newer computerized field polygraphs.

Introduction

Several methods are currently used by field polygraph examiners to measure electrodermal responses during polygraph examinations. Traditional analog polygraphs typically record skin resistance (SR) from large metal plates placed on the fingertips. These plates develop bias potentials, are subject to movement artifacts, and place high power dissipation requirements on individual sweat glands that can affect sweat gland activity. The recordings from analog instruments often show rapid drops in the baselines, which is a nuisance because the polygraph examiner must frequently re-center the recording pen. To avoid this, some polygraph manufacturers include a filter on their analog instruments that stabilizes the baseline, but it also alters the shape and amplitude of the examinee's SR response to test questions.

By the early 1970s, academic psychophysiologists had abandoned the recording of SR in favor of skin conductance (SC). The advantages in recording SC are that large bias potentials can be avoided, a low constant-voltage circuit (0.5V) is used that has little or no effect on the sweat glands, and the baseline is relatively stable, obviating the need for filtering. In addition, research indicates that SC is related linearly to the number of active sweat glands at the recording site, whereas SR is not (Venables & Christie, 1980). A disadvantage in using the low voltage SC circuit is that it requires the use of wet electrodes. The polygraph examiner must apply a small amount of paste to the electrode before placing it on the skin. The dry metal plates used with traditional analog polygraphs are more convenient because they do not require the use of a conductive electrode gel.

Honts and Barger (1990) compared SC and SR recorded from dry metal plate electrodes attached to four fingers of the same hand. SC was recorded on a traditional analog polygraph with a high, 2.2 V constant-voltage circuit manufactured by Lafayette Instruments (Lafayette, IN). They reported no difference between SC and SR in response amplitudes, although they did find that the SR channel required more pen adjustment than the SC channel during the test. Honts and Barger recommended SC because it requires less adjustment than SR, and it is more closely related to eccrine sweat gland activity.

The failure of Honts and Barger to observe a difference between the amplitudes of SC and SR responses was consistent with results from an earlier study by Boucsein and Hoffman (1979). In contrast to Honts and Barger, Boucsein and Hoffman used laboratory equipment that applied a constant-voltage (0.5V) or a constant current (10 μAmp/cm^2) to wet electrodes. Boucsein and Hoffman found only one difference; the recovery times of SC responses were shorter than were those of SR responses.

The present study evaluated one of several possible differences among methods for assessing subjects' electrodermal responses to test questions. Skin conductance was recorded with the constant-voltage circuit from wet electrodes. Continuous absolute measures of SC were digitized by a computer and subsequently were transformed to SR. Since resistance is the reciprocal of conductance, each calibrated value of conductance (in Siemens) was divided into 1.0 to obtain resistance in ohms. The amplitudes of SC and SR responses were then extracted from the respective response waveforms and compared. Consequently, the present study focused on only the effects of a nonlinear, but monotonic, transformation of SC to SR.

Method

Subjects

Four-hundred-and-seventeen adults were recruited from the general community by newspaper advertisements for a study that examined the effects of the demonstration test on the accuracy of probable lie and directed lie polygraph examinations (DoDPI97-P-0016). The advertisements offered $30 for two hours of participation and the opportunity to earn an additional $50 bonus. Of the 417 individuals, 81 were eliminated from the study. Thirty-three individuals assigned to the guilty condition (16%) declined to participate after they received instructions to commit a simulated theft. Eighteen individuals failed to follow instructions. For example, some individuals did not commit the theft, yet reported for their polygraph test. Others arrived too late or brought a child with them to the lab. Thirteen individuals were dismissed due to health problems, including reports of pain, less than 4 hours sleep, and high blood pressure. Nine individuals assigned to the guilty condition (5%) confessed during the pretest interview. Equipment problems and experimenter errors resulted in the loss of 5 other individuals. The remaining 168 innocent and 168 guilty participants were retained to fill the cells of the design matrix (described below).

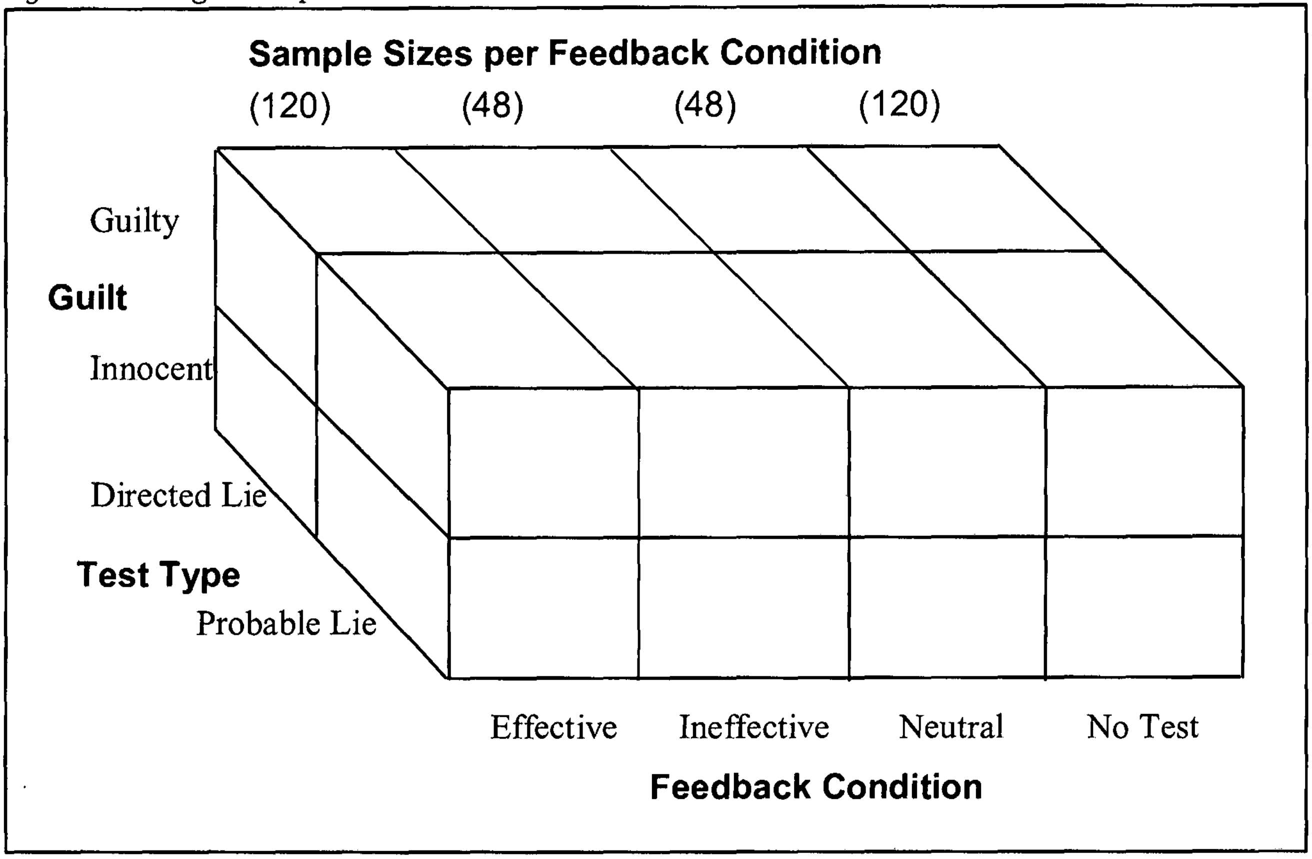
The mean age of participants was 30.7 years (SD = 11). Years of education ranged from 9 to 25 (M = 14.3, SD = 2.5). Most participants were Caucasian (87.5%) or Hispanic (5.7%) and either single (53.9%) or married (33.9%). A wide range of occupations was represented, the most common being student (17%), professional (11.9%), sales worker (9.2%), office worker (8.3%), service worker (8.3%), unemployed (7.7%), homemaker (7.7%), and laborer (7.5%).

Design

Participants were assigned randomly to one of 16 groups in a 2 X 2 X 4 factorial design, with equal numbers of male and female participants assigned to each cell. All factors except Sex are represented in the design illustrated in Figure 1.

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

Figure 1. Design of Experiment



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

Twelve participants were assigned to each of the four ineffective-feedback cells of the design matrix. Participants who received ineffective feedback were given a numbers test and told, regardless of the outcome, that they did not react appropriately to the chosen number and that they may not be suitable for examination using a polygraph.

Twelve participants were assigned to each of the four neutral-feedback cells of the matrix. Participants who received neutral feedback were given a numbers test and told that the numbers test would provide an opportunity for the participant to practice answering questions and for the polygraph examiner to adjust the instrument. Participants were given no information about the outcome of the numbers test.

Thirty participants were assigned to each of the four no-test control groups, as illustrated on the far right of Figure 1. Participants in the control groups were not given a numbers test. Otherwise, the pretest procedures were the same as those for all other participants.

To summarize, 120 participants were given the demonstration test and received feedback that the test was effective. Another 48 participants were given a demonstration test and received feedback that the test was ineffective. Another 48 participants were given a demonstration test and received neutral feedback. The remaining 120 participants were not given a demonstration test. Within each level of the Feedback factor, the design was balanced in terms of numbers of guilty and innocent male and female subjects who were given either probable-lie or directed-lie polygraph examinations.

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

Procedures

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

Callers who qualified and agreed to participate were given an appointment to report to a room in a building on the campus of the University of Utah. A map and reminder letter were mailed to participants who were scheduled more than a couple of days prior to their appointment. The participant also was called or sent an electronic mail message as a reminder the evening before their appointment.

An envelope addressed to the participant was taped to the door of the room to which the participant reported. Instructions within the envelope directed the participant to enter the room, close the door, read and sign an informed consent form, complete a brief questionnaire, and then play a cassette recorder that presented their instructions over earphones.

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

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

When the polygraph examiner went to the waiting room, he asked participants to use the restroom and wash their hands. When they returned from the restroom, they were escorted to the laboratory and asked to sit in the examinee's chair. The session was video and audio taped using a camera mounted high on the wall in front of the participant.

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

After the demonstration test, the examiner reviewed the appropriate set of questions with the participant. The test questions for participants assigned to the probable-lie condition were as follows:

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

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

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

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

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

(Neutral) 6. Is today ___?

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

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

(Neutral) 9. Is your first name ___?

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

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

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

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

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

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

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

Participants were told that it was very important that they appear to be lying to the directed-lie questions. The examiner told the participant that he would not want to make a mistake and conclude that they had lied if they were in fact telling the truth, simply because they did not appear to be lying on these questions.

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

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

At the conclusion of the test, the sensors were removed, and the subject was asked to complete post-test questionnaires. After minor editing of obvious movement artifacts from the physiological data, the probability that the participant was truthful was then determined using computer algorithms described elsewhere (Kircher & Raskin, 1988). If the probability of truthfulness exceeded 0.70, the participant was diagnosed as truthful and paid $80, $30 for their time and a $50 bonus. If the probability of truthfulness was less than 0.30, the participant was diagnosed as deceptive. If the probability of truthfulness was between 0.30 and 0.70, the test was considered inconclusive. If the outcome was deceptive or inconclusive, the participant was paid only $30. The participant was then debriefed and released.

Apparatus

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

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

Skin conductance was obtained by applying a constant voltage of 0.5V to two Beckman 10mm Ag-AgCl electrodes filled with .05 M NaCl in a Unibase medium. The electrodes were attached with double-sided-adhesive collars to the middle phalanx of the ring ànd smallest fingers of the left hand. The signal was recorded DC-coupled with a 2-pole, low-pass filter, fc = 6 Hz.

Respiration was recorded from two Hg strain gauges secured with Velcro straps around the upper chest and abdomen, just below the ribcage. The strain gauge changed in resistance as the subject breathed. Resistance changes were recorded DC-coupled with a 2-pole, low-pass filter, f_c = 8.8 Hz.

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

The procedures for measuring finger pulse, electrocardiogram, skin potential, and blood pressure are described in detail in another report (Kircher, Packard, Bell, & Bernhardt, 2001).

The 1000 Hz samples for each channel were reduced prior to storing them on the hard disk by averaging successive sample points. Electrodermal and respiration channels were stored at 10 Hz. Cardiograph signals were stored at 100 Hz.

Calibration Procedures

To test for differences between skin conductance and skin resistance measures, it was necessary to convert the raw data provided by the analog-to-digital converter to absolute units of skin conductance and, subsequently, resistance. To measure skin conductance, a separate multiple regression equation was developed for each of six possible gain settings on the PDAS. Each equation predicted known conductances from the offset on the front panel, internal PDAS digital-to-analog (DAC) offset settings, and observed analog-to-digital converter values. The conductance values used to calibrate the instrument ranged from 1 μSiemen to 50 μSiemens. External (front panel) and internal (DAC) offsets were also systematically varied to ensure that the resulting equation would work for any combination of gain and offset settings. Each equation accounted for over 99.8% of the variance in known inputs.

Since resistance (R) is the reciprocal of conductance (G), skin resistance was obtained by inverting the scaled skin conductance signal prior to extracting measurements of response amplitude, i.e., $R = 1 / G$.

Feature Extraction

Feature extraction was accomplished with the CPS-LAB computer program. The features extracted from each physiological channel were those that have been found in previous investigations to be optimal for the prediction of deception in a laboratory mock crime study.

Skin Conductance and Skin Resistance.

For each comparison and relevant question, the peak amplitude of SC and SR was extracted from a 20-second interval that began at question onset. Although the dependent variables of greatest interest were the amplitudes of SC and SR responses, additional response features were extracted to test if there was any systematic difference between SC and SR responses in terms of their overall diagnosticity. Figure 2 illustrates the measurement of response amplitude and selected additional features. Full recovery time, duration to full recovery, full recovery rate, and area are not illustrated in Figure 2, but they were also measured. The point of full recovery occurred when the recovery limb of the response reached the level at response onset. The response shown in Figure 2 did not recover to the level at response onset before the end of the 20-second scoring window. When the response did not recover completely, the point of full recovery was taken as the end of the 20-second scoring window.

Cardiograph.

The times and levels of systolic and diastolic points were identified in the cardiograph signal and used to create second-by-second systolic and diastolic response curves (Kircher & Raskin, 1988). A trend line was computed by calculating the mean of the systolic and diastolic points for each second. The maximum increase in the trend line (response amplitude) was extracted from the 20-second interval that followed question onset.

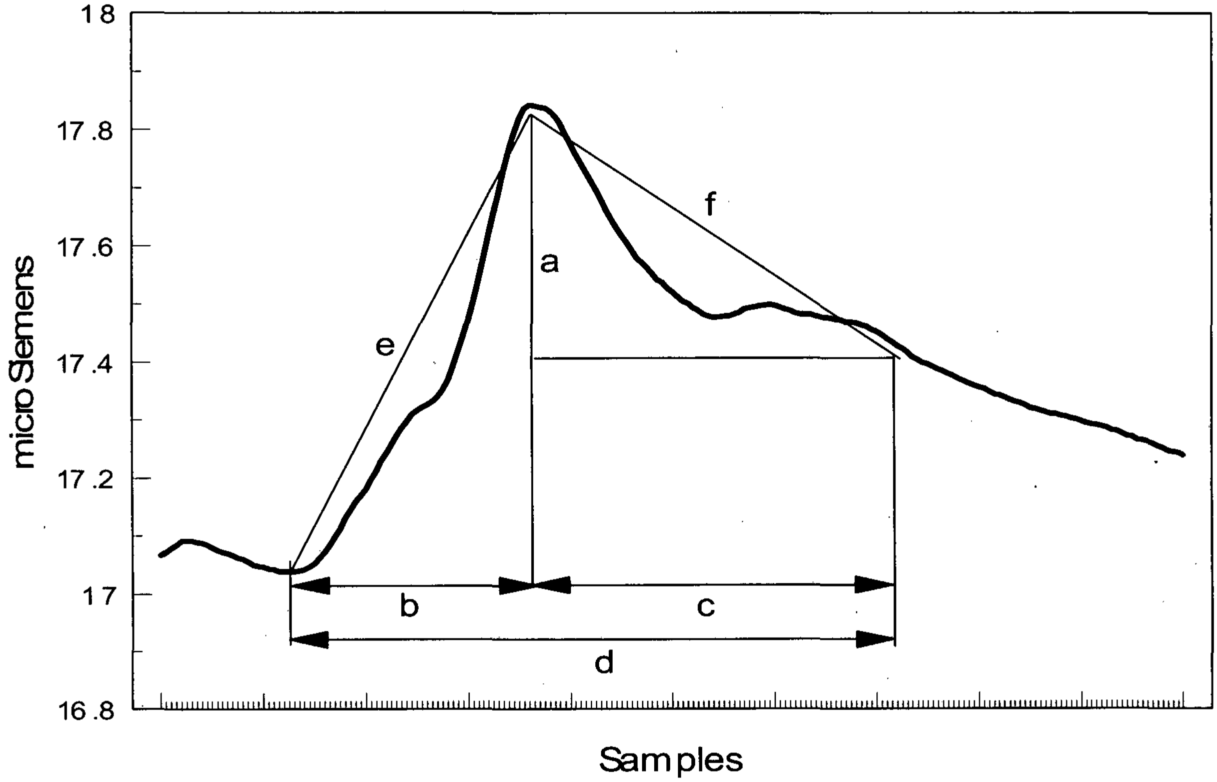
Thoracic and Abdominal Respiration.

Thoracic and abdominal respiration excursion was measured for a period of 10 seconds following question onset. Since each respiration channel was stored at 10 Hz, 100 samples showed the respiration activity for the 10-second interval. Respiration excursion was the sum of 99 absolute differences between adjacent samples.

Feature Standardization

The present analyses were limited to the first three charts of physiological data. Each chart contained three comparison questions and three relevant questions. The six questions on each of the three charts provided 18 repeated measures of each physiological feature. The set of 18 measurements of a given feature were transformed to standard scores (z-scores) within the subject. A mean z-score was calculated for the comparison questions and another mean z-score was calculated for the relevant questions. The difference between the means for comparison and relevant questions

Figure 2. Feature Extraction



a. Amplitude
a. Rise time
b. Half-recovery time
c. Duration to half recovery (b+c)
d. Rise rate (a/b)
e. Half-recovery rate (a/2 /c)

provided an overall index of differential reactivity to comparison and relevant questions.

For SC amplitude, SR amplitude, and cardiograph amplitude, a large measured value was interpreted as a strong physiological reaction to the question. For respiration excursion, a relatively small measured response was indicative of strong respiration suppression. To establish a common direction for predicted effects, the sign of the difference between comparison and relevant questions was reversed for respiration excursion. Thus, for all measures, stronger reactions to comparison questions resulted in positive difference scores, and stronger reactions to relevant questions resulted in negative difference scores. A single composite measure of differential respiration suppression was then obtained by computing the mean of the difference scores for thoracic and abdominal respiration excursion.

Results

Preliminary Analyses

Treatment-Related Attrition

Thirty-three individuals assigned to the guilty condition (15%) refused to participate after they had received their tape-recorded instructions, whereas none of the innocent subjects declined to participate. Consequently, individuals who agreed to commit the mock crime may have been sampled from a population that differed in certain respects from the more general population from which innocent participants were drawn. For example, participants in the guilty condition on average may have been older or less anxious than were participants in the innocent condition. Preliminary tests were conducted to explore the possibility that guilty and innocent participants differed on measures of marital status, ethnicity, occupation, age, education, hours of sleep, the Marlowe-Crowne scale (Crown & Marlowe, 1964), Rotter Trust Scale (Rotter, 1967), and two anxiety scales (Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983). The guilty and innocent participants who completed the experiment did not differ significantly on any of the demographic or personality measures.

Effects of Guilt, Feedback, and Test Type

A split-plot ANOVA was performed to test if differences between SC amplitude and SR amplitude varied across the Guilt, Feedback, and Test Type facets of the design. Method of measurement was the within-subjects factor with two levels (SC amplitude versus SR amplitude). Guilt, Feedback (effective, neutral, ineffective, and no test), and Test Type (PL versus DL) were between-group factors. If differences between SC amplitude and SR amplitude varied over levels of Feedback or Test Type, it might not be possible to conclude that one electrodermal measure was generally superior to the other.

The results of the split-plot ANOVA revealed no main effect of Method nor was there any significant 2-, 3-, or 4-way interaction of Method with Guilt, Feedback, or Test Type. The lack of a significant Guilt X Method interaction suggests that SC amplitude and SR amplitude were similarly useful for discriminating between guilty and innocent participants when the data were pooled across Test Types and Feedback conditions. The lack of any higher-order interactions that included Guilt and Method as factors suggests that the *difference* between SC amplitude and SR amplitude in their ability to discriminate between guilty and innocent subjects (Guilt X Method interaction) did not depend on the Test Type or Feedback.

Univariate Analyses

The objective of the present study was to determine if there is an advantage in using SC or SR for the detection of deception. To that end, the preliminary analyses indicated that it is reasonable to pool the data across PL and DL tests and across Feedback conditions and to capitalize on the large sample size. Therefore, comparisons based on all 336 participants are presented below. However, the preliminary ANOVA also revealed small but significant Guilt X Test Type, $F(1, 320) = 5.49$, $\eta^2 = .02$, and Guilt X Feedback X Test Type interactions effects, $F(3, 320) = 2.73$, $\eta^2 = .03$, on both electrodermal measures. Since only the effective feedback condition was representative of current field practice, results also are reported separately for the PL and DL effective feedback conditions.

Table 1 presents the means of the differences between standardized responses to comparison and relevant questions for the SC amplitude and SR amplitude measures. The means are presented separately for the guilty and innocent participants in the entire sample and for the PL and DL effective feedback subsamples. Table 1 also shows the point-biserial correlation between the index of differential reactivity and a dichotomous criterion variable that distinguished between the guilty (coded 0) and innocent participants (coded 1) in a sample. The point-biserial correlation is a measure of diagnostic validity; it indicates the extent to which the physiological measure discriminated between the innocent and guilty groups.

Overall, the validity coefficients for SC and SR were lower in the entire sample (M_r = .61) than in either of the effective feedback subsamples (Probable-Lie M_r = .72; Directed Lie M_r = .64). ANOVAs revealed significant Guilt X Feedback interactions for SC ($F(1, 328) = 4.59$, $p < .05$) and SR ($F(1, 328) = 4.04$, $p < .05$) when the effective feedback conditions jointly were compared to the groups that did not receive effective feedback. These findings suggest that a preliminary demonstration test and effective feedback enhance the accuracy of polygraph tests. These findings are discussed in detail in another report (Kircher, Packard, Bell, & Bernhardt, 2001. There was

Table 1. *Means (Standard Deviations) and Point-Biserial Correlations (r_{pb}) for SC Amplitude and SR Amplitude*

	Innocent		**Guilty**		**r_{pb}**
Entire Sample (N=336)					
SC Amplitude	.46	(.62)	-.50	(.64)	.61**
SR Amplitude	.44	(.62)	-.52	(.64)	.61**
Probable-Lie Effective Feedback (n=60)					
SC Amplitude	.42	(.62)	-.84	(.60)	.72**
SR Amplitude	.40	(.62)	-.74	(.62)	.72**
Directed-Lie Effective Feedback (n=60)					
SC Amplitude	.42	(.70)	-.62	(.58)	.63**
SR Amplitude	.40	(.66)	-.64	(.62)	.64**

** $p < .01$

a tendency for the electrodermal measures to be more diagnostic for effective feedback participants who received PL tests (M_r = .72) than DL tests (M_r = .64), but the difference was not significant for SC amplitude or for SR amplitude. Of greater interest in the present study were comparisons of SC and SR measures within treatment conditions. The validity coefficients for SC amplitude and SR amplitude were virtually identical for the entire sample and within the PL and DL effective feedback conditions. A separate z-test for correlated correlations (McNemar, 1969) was performed for each sample. The z-tests confirmed that there were no significant differences between the validity coefficients for SC and SR measures for the entire sample (.61 versus .61), the PL effective feedback group (.72 versus .72), or the DL effective feedback group (.63 versus .64).

Numerical evaluations of electrodermal responses depend primarily on measures of response amplitude (Bell et al., 1999; Swinford, 1999), and the only feature extracted from the electrodermal response by our CPS computer program is peak amplitude. However, other aspects of the electrodermal response, such as duration and number of responses, are used by numerical evaluators and might be used by computer programs in the present or future. Table 2 presents the validity coefficients for selected features of the SC and SR waveforms, including the amplitude measures presented previously. Since the SC and SR waveforms are monotonically related, there was always an equal number of SC and SR responses. Although there was no possibility that the number of SC and SR responses would differ, the point-biserial correlations for number of responses are reported for completeness.

Table 2. *Point-biserial Correlations with the Guilt/Innocence Criterion for Features Extracted from SC and SR Waveforms*

	Entire Sample (N=336)		*Probable-lie Effective Feedback (n=60)*		*Directed Lie Effective Feedback (n=60)*	
Feature	**SC**	**SR**	**SC**	**SR**	**SC**	**SR**
Amplitude	.61	.61	.72	.72	.63	.64
Rise time	.23	.23	.25	.25	.37	.37
Half-recovery time	.48	.48	.55	.56	.43	.44
Full-recovery time	.57	.57	.64	.64	.61	.61
Duration to half-recovery	.37	.38	.40	.41	.45	.45
Duration to full-recovery	.57	.57	.65	.65	.62	.61
Number of responses	.25	.25	.39	.39	.30	.30
Area to half-recovery	.55	.55	.68	.66	.54	.53
Area to full-recovery	.58	.58	.71	.69	.56	.56
Rise rate	.57	.58	.67	.67	.63	.63
Recovery rate to half-rec	.43	.43	.57	.58	.52	.52
Recovery rate to full-rec	.36	.36	.55	.55	.35	.35

Note: All of the correlation coefficients in Table 2 were statistically greater than zero, $p < .05$.

The results in Table 2 are consistent with those obtained for SC amplitude and SR amplitude. There was virtually no difference in diagnostic validity of various aspects of the SC and SR response waveforms.

Boucsein and Hoffman (1979) found no difference between the amplitudes of SC and SR responses, but they did report that the recovery times of SC responses were shorter than were those of SR responses. To assess the reliability of that finding, the mean of raw measurements of SC half recovery time across all comparison and relevant questions was obtained for each participant. Another mean was obtained for SR half recovery times. Method of measurement (SC versus SR) was treated as a repeated measure in an ANOVA with Guilt, Test Type, and Feedback as between-group factors. Consistent with the results reported by Boucsein and Hoffman, SC half recovery times (M = 3.77 sec) were significantly shorter than the half recovery times of SR responses (M = 4.08 sec), $F(1, 320) = 373.7$, $p < .01$.

Multivariate Analyses

It is conceivable that one variable may be correlated with the criterion as highly as or even less highly than another, and yet it produces higher hit rates when combined with other physiological measures. To evaluate this possibility, a discriminant function was created to classify cases that included SC amplitude along with respiration excursion and cardiograph baseline increases. Another discriminant function was created that included SR amplitude as well as the respiration and cardiograph measures. Cases were classified as truthful if the probability of truthfulness exceeded 0.70, deceptive if the probability of truthfulness was less than 0.30, and inconclusive if the probability was between 0.70 and 0.30. Percent correct, wrong, and inconclusive (?) that resulted from this decision rule are presented in Table 3.

Table 3. *Percent Outcomes for Discriminant Functions that Combine SC Amplitude or SR Amplitude with Cardiograph and Respiration Measures*

	Innocent				**Guilty**			
	Correct	Wrong	?	Correct Decisions	Correct	Wrong	?	Correct Decisions
Entire Sample (N=336)								
SC Amplitude	63.7	9.5	26.8	87.0	63.1	10.1	26.8	86.2
SR Amplitude	62.5	9.5	28.0	86.8	64.9	10.7	24.4	85.8
Probable-Lie Effective Feedback (n=60)								
SC Amplitude	90.0	6.7	3.3	93.1	86.7	6.7	6.7	92.9
SR Amplitude	90.0	3.3	6.7	96.4	83.3	6.7	10.0	92.6
Directed-Lie Effective Feedback (n=60)								
SC Amplitude	70.0	16.7	13.3	80.8	73.3	10.0	16.7	88.0
SR Amplitude	76.7	16.7	6.7	82.1	70.0	10.0	20.0	87.6

In the entire sample, and within each effective feedback condition, there was little difference between outcomes of discriminant functions that included SC or SR responses. A post hoc examination of Table 3 suggested that the inconclusive rate appeared higher in the entire sample than in either of the effective feedback subsamples. To test this hypothesis, subjects who received effective feedback were dropped from the entire sample, and frequencies of correct decision, wrong decision, and inconclusive were calculated for the remaining three groups. Another group was formed by pooling the results from the PL and DL subjects who received effective feedback. A chi-square test was then conducted to test if the distribution of outcomes (correct, wrong, and inconclusive) depended on whether or not the subject had received effective feedback. The test confirmed that the presentation of effective feedback affected outcomes, $X^2(2) = 27.8$, $p < .01$. The effect was due primarily to the relatively low number of inconclusive cases in the group that received effective feedback. Another chi-square test compared outcomes from the PL and DL effective feedback groups. Although there appeared to be some advantage in using the PL test, the difference was not statistically significant, $X^2(2) = 5.4$, $p < .07$.

Discussion

The results of our comparisons of SC and SR are consistent with those reported by Boucsein and Hoffman (1979) and by Honts and Barger (1990). We observed no differences between SC amplitude and SR amplitude measurements across a wide range of treatment conditions. Boucsein and Hoffman also reported that the recovery time of SC responses was shorter than the recovery time of SR responses. We replicated that finding as well. However, within-subject comparisons of SC and SR recovery times revealed no differences in their ability to distinguish between truthful and deceptive individuals.

In the present study, SC recordings were transformed mathematically to SR prior to extracting features from them. Although this transformation was nonlinear, it had no discernable effect on any of the 12 response characteristics measured in the present study. Visual comparisons of individual SC and SR responses by several participants suggested that within the range of measurements for an individual, the inverse transformation from SC to SR was essentially linear. The greatest observed difference in the shape of the SC and SR waveforms occurred when the basal level of SC was low (e.g., 1μS) and the reactions were large relative to the basal level (e.g., 0.5 μS). Even then, the transformation produced an SR waveform that appeared very similar to the SC waveform on most dimensions. The results of visual inspection and computer analysis were consistent; SC and SR response waveforms were virtually identical.

The consequences of failing a polygraph test administered during an actual criminal investigation are usually much greater than those associated with failing a test in a laboratory experiment. On average, tonic levels of arousal may be greater in the field than in the laboratory. Given that SC and SR responses appear more similar at elevated levels of tonic activity, it seems unlikely that significant differences between SC and SR measures would emerge in the field. Nevertheless, it should be noted that this was a mock crime experiment and the present findings may not generalize to the field.

The present study evaluated only one of several possible differences in methods for measuring participants' electrodermal responses to test questions. SR was based on measurements obtained with a constant low-voltage source, but it is usually obtained from a constant current source. The design of the present study did not permit a test for possible differences in recording techniques. In terms of decision accuracy, prior research suggests there would be no particular advantage in choosing a constant-voltage over a constant-current circuit for polygraph testing (Boucsein & Hoffman, 1979; Honts & Barger, 1990). However, we agree with Honts and Barger (1990) that the constant-voltage SC circuit is preferred. We agree because it produces measures that are related linearly to the number of active eccrine sweat glands, it produces a stable baseline that does not require a high pass filter, and it is consistent with accepted scientific practice. Boucsein and Hoffman (1979) used laboratory equipment with wet nonpolarizing electrodes, whereas Honts and Barger (1990) used an analog field polygraph with dry metal plate electrodes. Since neither study compared laboratory instrumentation to field polygraphs, we do not yet know if laboratory equipment and techniques yield electrodermal measures that are more diagnostic of deception than those from traditional analog polygraph instruments.

Another unanswered question concerns the differences among computerized polygraph systems in their methods for recording electrodermal activity. The Computerized Polygraph System (CPS; Stoelting Company, Wood Dale, IL), the Axciton system (Axciton Systems, Houston, TX), and the Lafayette system (Lafayette Instruments, Lafayette, IN) are currently used by field polygraph examiners. Only the CPS system records SC from wet electrodes with a constant-voltage circuit and meets guidelines for recording electrodermal activity established by the scientific community (Fowles, Christie, Edelberg, Grings, Lykken, & Venables, 1981). The other systems use traditional dry metal plates as electrodes. However, in contrast to traditional analog instruments, the Axciton and Lafayette computerized polygraphs do not measure SR. In his tests of the three computerized polygraph systems, Cestaro (1998) found that the signals generated by Axciton and Lafayette computer systems did not accurately reproduce known changes in resistance or conductance. Although there are limitations to the traditional methods for recording SR, at least there is a more or less direct (monotonic) relationship between SR and the activity of the eccrine sweat glands (Venables & Christie, 1980). In light of Cestaro's findings, the same cannot be said of the electrodermal signals generated by the Axciton and Lafayette computerized polygraphs. Therefore, it is also important to determine if the electrodermal measures from laboratory-grade polygraph instruments, such as CPS, are more useful for detecting deception than those provided by other computerized polygraph instruments.

References

Boucsein, W. & Hoffman, G. (1979). A direct comparison of skin conductance and skin resistance methods. *Psychophysiology, 16*, 66-70.

Cestaro, V. (1998). Memorandum for Record: Laboratory tests performed on the electrodermal activity (EDA) channels of various polygraph instruments. Report on project DoDPI98-P-0003 to the U. S. Department of Defense Polygraph Institute, Ft. McCellan, AL.

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

Honts, C. R. & Barger, S. D. (1990). A comparison of the relative utility of skin conductance and skin resistance couplers for the measurement of electrodermal activity. *Polygraph, 19,* 199-207.

Fowles, D. C., Christie, M. J., Edelberg, R., Grings,W. W., Lykken, D. T., & Venables, P. H. (1981). Publication recommendations for electrodermal measurements. *Psychophysiology, 16,* 66-70.

Kircher, J. C., Packard, T., Bell, B. G. & Bernhardt, P. C., (2001). Effects of Prior Demonstrations of Polygraph Accuracy on Outcomes of Probable-Lie and Directed-lie Polygraph Tests. Final report to the U. S. Department of Defense Polygraph Institute, Ft. Jackson, SC. Salt Lake City: University of Utah, Department of Educational Psychology.

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

Kircher, J. C., Woltz, D. J., Bell, B. G., & Bernhardt, P. C. (1998). Effects of audiovisual presentations of test questions during relevant-irrelevant polygraph examinations and new measures. Final Report to the U. S. Central Intelligence Agency. Salt Lake City: University of Utah, Department of Educational Psychology.

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

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

Venables, P.H. & Christie, M.J. (1980). Electrodermal activity. In I. Martin & P.H. Venables (Eds.) *Techniques in Psychophysiology.* Chichester: Wiley.

Does the Confession Criterion in Case Selection Inflate Polygraph Accuracy Estimates?

Donald J. Krapohl, Kendall W. Shull, and Andrew H. Ryan

Abstract

Many polygraph field studies have relied on confessions as verification of ground truth, a criterion that some critics argue creates an overestimation of polygraph accuracy. This is because there is a relationship between polygraph results and the likelihood that a suspect will confess. Confessions come from interrogations, which follow failed polygraphs. If a guilty person fails the polygraph, an interrogation is initiated, which might yield a confession. If a guilty person passes the polygraph, there is no interrogation, no confession, and little chance the polygraph error will be uncovered. This would suggest that among guilty suspects, there could be qualitative group differences between confession and nonconfession cases. The biasing effect of this confession criterion has not yet been resolved. In this study, a comprehensive sample of field polygraph cases from a large U.S. government polygraph program was examined to uncover differences in the polygraph detectibility of guilty confessing suspects, and guilty suspects who did not confess but were caught by other means. The present data failed to find any differences in the groups. This manuscript does, however, correct errors published elsewhere regarding law enforcement polygraph and investigative practices in the field.

Introduction

Among forensic disciplines, none is as controversial as using the polygraph to detect deception. The use of the polygraph to uncover criminal and security-related behaviors now spans seven decades and has been the center of heated debate for virtually the entire period. There are many facets to the debate, but the most frequent issue centers on the accuracy of the comparison question technique, the most common polygraph technique in the field. Critics have charged that the comparison question technique (formerly known as the control question technique) lacks validity and argue that the empirical evidence is, at best, incomplete. Proponents agree that more research is needed, but argue that the preponderance of the available field data points to an accuracy of about 90 percent.

Critics are not as comfortable with the available field studies as are the proponents. It is well known that the method in which cases are selected for a study affects the outcome of the study and that some methods are better than others. Polygraph critics contend that existing research supporting polygraphy has systematically stacked the deck in favor of higher accuracy. The biggest culprit, according to some (Ben-Shakhar et al. 1982; Lykken 1998; Patrick and Iacono 1991), is the confession criterion. The confession criterion allows polygraph cases to be selected for research based on the confession of the examinee. This use of the confession criterion may bias the types of cases used in a field study. The confession criterion could inflate accuracy estimates in detecting deception by the way comparison question technique field studies are typically conducted. This is explained in the following paragraphs.

To test the efficacy of the comparison question technique, it is necessary to have confirmed cases, that is, polygraph recordings from a group of examinees for whom ground truth has been unquestionably established.

Ground truth is easy to determine in laboratory studies because experimenters assign examinees their roles of guilt or innocence. In the field, on the other hand, examinees arrive for polygraph appointments

[1] Reprinted with permission from Forensic Science Communications. July 2002, Volume 4, Number 3.

with self-assigned roles, usually not known to anyone except themselves and their collaborators. Therefore, experimenters must resort to other means to determine ground truth in field studies.

In polygraph field research, the use of the confession criterion is fairly common. The confessions of examinees are the most readily available confirmations, but this is where the problem begins. Guilty examinees typically do not spontaneously confess their crimes or deceptions to polygraph examiners or investigators. They are far more likely to acknowledge their acts during an interrogation. However, in standard polygraph practice, the only examinees who are interrogated are the ones who have failed the polygraph examination. If a guilty person manages to pass the examination, there probably would be no confession because there would have been no interrogation. Therefore, data sets consisting only of confession-confirmed cases might contain merely those where deception was most apparent in the test charts. Cases where the polygraph was fooled would not be found in the sets. As Iacono (1991) points out:

> "Because polygraphers seldom discover ground truth except as a consequence of post-test confessions, and because diagnoses evaluated in this way are almost invariably verified as correct, the typical experienced examiner will accumulate a personal record of almost unblemished accuracy (p. 202)."

A similar problem exists with misdiagnosed innocent examinees. If an innocent examinee fails a polygraph examination, he or she almost never confesses, even when interrogated. Unless evidence surfaces that someone else was actually guilty, the case remains unconfirmed and, therefore, would not be selected for accuracy studies. Iacono (1991) adds that cases are closed when an examinee has a deceptive outcome on the polygraph, thereby cutting off the possibility of the discovery of disconfirming information. This policy would reduce the likelihood of an agency ever uncovering the true guilty party and discovering the polygraph error.

Horvath (1977) was the first to investigate the possible relationship between confessions and polygraph accuracy. He drew a sampling of verified and unverified polygraph cases from the files of criminal suspects at a large police agency. He used an equal number of deceptive and nondeceptive cases from the verified and unverified categories, with a total of 112 cases used in the study. The cases were selected randomly to fill the cells, and the criterion for verification was the confession of an examinee. This inculpated the examinee and exculpated others being polygraphed for the same crime. The cases were subjected to blind analysis by ten field examiners who worked in law enforcement. Horvath did not find any differences in the scorers' decisions with verified and unverified cases. These results led him to cautiously conclude that confession cases did not enjoy better discernment by polygraph examiners, although he recommended further investigation.

Raskin et al. (1988) evaluated all of the U.S. Secret Service polygraph cases for a 2½ - year period and found 76 cases where ground truth was established independently of the polygraph results. Raskin used a two-step process in case confirmation where there was a confession that inculpated or exculpated the examinee, and there was independent physical evidence consistent with the confession. To investigate the possible effects of the confession criterion, Raskin added 20 unconfirmed cases to the set. The 96 cases were then scored manually by U.S. Secret Service polygraph examiners who did not know the ground truth for any of the cases. Raskin reported that the average polygraph scores of confession-confirmed guilty cases along with unconfirmed guilty cases were different by approximately 20 percent. In cases where examinees confessed, the scores were an average of 20 percent more in the deceptive direction than in cases that were decided as deceptive but unconfirmed. At face value, these findings supported the argument that the confession criterion yields inflated accuracy estimates because the confession cases appeared easier to diagnose. However, the Raskin conclusions were mitigated by the findings that the unconfirmed guilty cases had scores 63 percent beyond the threshold needed to make a conclusive decision. In other words, the effect was statistically significant, but effectively trivial.

In their experimental design, Raskin et al. (1988) attempted to control the sampling bias among the innocent cases by requiring that each confirmed innocent case be part of a multiple-suspect investigation in which the culprit was found or that the crime be determined not to have taken place. In that way, any false-positive outcomes could be discovered without biasing the sample. This study has adopted Raskin's safeguard.

It is prudent to agree with Iacono (2000) who suggested that this safeguard, by itself, might still have two possible weaknesses. First, if a polygrapher knew the outcome of other suspects' tests, it is not unreasonable that this knowledge could influence how subsequent examinations are interpreted. In the perfect field study, all of the suspects would be polygraphed separately by polygraphers who did not know the number of suspects or the outcomes of the other polygraph examinations. In that way, the polygraph decisions could not be affected by examiner expectancies, one source of variability shown to influence polygraph scoring (Elaad et al. 1994). To control this potential scoring bias in the present study, an automated analysis method was applied that relies on measurements of tracing features rather than on the semiobjective scoring system used in the field. This approach, described later, avoids the confounding influence of examiner expectancies on chart interpretation.

A second potential source of selection bias of innocent cases, Iacono (2000) suggests, is that polygraphers who believe so strongly in their results do not usually test any further suspects in a case once one has failed the polygraph examination. If the failed suspect is actually innocent, subsequent investigative resources can be misdirected, resolution of the case can become more difficult, and the polygraph error can become less likely to be discovered. However, when the polygrapher correctly identifies a suspect, the decisions of nondeception would be confirmed for previous cases. Therefore, when the testing examiner makes the right decision, confirmation is more likely to arise.

There are two assumptions in Iacono's (2000) hypothesis that bear closer scrutiny. First is the assertion that polygraphers believe in their exams so strongly that they usually stop testing other suspects once one has failed. It should be noted that polygraphers in the U.S. federal government are not empowered to choose whom to polygraph or not to polygraph. These decisions rest in the hands of investigators, managers, and prosecutors whose distance from the polygraph makes them less vulnerable to the errors of such blind acceptance. It is also worthy of note that it would be quite uncommon for any state or local law enforcement agency in the United States to delegate the decision of whom to polygraph to its staff polygraphers. Thus, Iacono's (2000) assumption in this regard does not apply to examinations conducted in the United States.

The notion that polygraphing stops after an examinee is found deceptive, regardless of who decides whether or not to continue, also communicates an incomplete understanding of law enforcement investigative practices. At the heart is the misapprehension that law enforcement agencies act as though all crimes have a single culprit, that there are no coconspirators or partners that might also be on the list of suspects. In the real world, the decision to stop polygraphing depends on whether investigators are satisfied that all of the perpetrators have been identified, not on whether the polygrapher caught one. Iacono's (2000) assumption is incorrect on this aspect, as well.

Patrick and Iacono (1991) also examined the sampling bias issue in a field study carried out on police cases from Vancouver, British Columbia. Beginning with 402 possible cases, they pared it to a sample of 89 cases where ground truth was verified to what Patrick and Iacono characterized as "maximum certainty"-37 were innocent, and 52 were guilty. Among the 52 guilty, according to the Patrick and Iacono criteria, no false negatives were found in their exhaustive review of the evidence. They found that ground truth as determined by examiner-verified cases did not match those of their own strict confirmation criteria. Examiners were far more lenient in their judgments for confirmation of their own work. For example, an examinee was called deceptive on his polygraph examination, and during the post-test interrogation he admitted to committing a crime, though not the specific crime covered in the relevant test questions. The examiner still labeled the case as confirmed. Patrick and Iacono asserted that many comparison question technique field studies are based on just these types of data in which accuracy in detection of deception is inflated by the generous criteria that polygraph examiners afford themselves. Patrick and Iacono overcame this shortcoming through a more rigorous verification process. In addition to the blind scoring of the charts to remove extra polygraphic sources of information, they found that the polygraph decisions were 98 percent correct with guilty examinees, even with their criteria. Patrick and Iacono also reported that post-test confessions were related to highly negative (deceptive) scores. Correct classification of the innocent cases in the Patrick and Iacono study was near chance levels with blind scorers. Though accuracy was far lower than that achieved by the original examiners, the researchers proposed that the blind scoring results of those 37 cases were representative of polygraphy in the field.

Honts (1996) conducted a partial replication of the Patrick and Iacono (1991) study with a smaller data set but developed an innovative approach to test for the biasing effects of the confession criterion. Honts devised a scaling system that quantified the level of confirmation for the cases. The assumptions of the confession criterion bias would lead to the expectation that polygraph scores (and hence, decisions) would be related to the degree in which the criminal cases produced independent evidence. Honts' results suggested that there was no effect on polygraph scores for the level of confirmation of ground truth; there was no meaningful effect for the confession criterion. His data also confirmed the high accuracy of guilty cases that Patrick and Iacono (1991) reported but found much better accuracy with the innocent cases than those from the Patrick and Iacono sample.

Honts suggested that the Patrick and Iacono (1991) study may have been an outlier because other similar studies (Honts and Raskin 1988; Raskin et al. 1988) found comparable accuracy for guilty and innocent examinees. As one explanation for the discrepant findings, Honts suggested that criterion contamination may have been an issue in the Patrick and Iacono (1991) study, a factor Honts stated had been controlled in the other research. In polygraph studies, criterion contamination can take place when an examinee's intention to deceive is captured by the examination, though not specifically to the relevant question at hand. Honts used the example taken from the Patrick and Iacono study where one of the polygraphers shared the following details: a suspect had been given a relevant question-"Did you steal the diamond ring?" The examinee was found deceptive on the polygraph examination and was confronted. He denied that he had stolen the ring but admitted that his brother had. The examinee's part in the crime was only that he had sold the stolen ring. According to Honts (1996), Patrick and Iacono reported this as a false-positive error because the examinee was called deceptive, even though he was not guilty of the specific relevant question. Honts argued otherwise, pointing to the examinee's intention to deceive about the ring theft. The issue is the subject of contentious debate even today. Readers wanting the full flavor are directed to the relevant chapters on polygraphy in Faigman et al. (1997).

The unbalanced accuracies in the Patrick and Iacono (1991) study may also have been the consequence of how the polygraph was applied to criminal investigation by that polygraph agency. In some settings the polygraph is used more generally to simply determine who should remain on the list of suspects. In other words, the appearance of unfair polygraph outcomes in Vancouver may have been that the decision rules were set so that no guilty examinees would escape, but that some percentage of innocent examinees would pass through. In the end, this method concentrates the suspect pool so that investigative resources can be more wisely invested. Because the polygraph was not used to incarcerate or convict the suspects, there was a relatively small cost to a false-positive outcome that might spring from biased decision rules. Those innocent examinees were on the suspect list before being polygraphed, and the polygraph examination merely failed to remove them from that list. In view of the potential harm to the community that can arise from a false-negative decision, especially when speaking of violent offenders, it may be that some police agencies adjust the decision rules to ensure those examinees are correctly classified, even when it means retaining some innocent examinees on the list. However, very different decision rules may be appropriate for other circumstances where there are more dire consequences for false-positive outcomes. For example, a far more balanced approach is warranted when polygraph evidence is used in courts of law.

Getting around suspected sampling problems has not been easy, and to date no mutually satisfying solution has been reported in the literature. Lykken (1998) proposed a novel approach to the investigation of accuracy of field polygraphy. He suggested that the FBI could use its own polygraph examiner staff, employing the polygraph in its usual manner, but to set aside the results of the examinations, and take no action; that is, no interrogations. At some later date, a panel would try to verify ground truth from all available evidence and compare it to blind scorings of the polygraph data.

Though it is interesting from an academic viewpoint and would help answer the question of polygraph accuracy, Lykken's is not a practical proposal because the FBI is not likely to be persuaded to ignore one of its forensic tools when there are serious crimes to solve. A less intrusive approach is proposed here, at least with regard to deceptive examinees, and it begins with this assumption: if the confession criterion causes a bias in the sampling of field cases, there should be qualitative group differences in scores and decisions between guilty confessors and guilty nonconfessors. Guilty confessors are those who would collectively have their deceptions more apparent on the polygraph charts. This would be consistent with Patrick and Iacono's (1991) report that confessions corresponded with more deceptive scores in most cases. If there are no significant differences in polygraph scores or results between confessors and nonconfessors, the impact of the confession criterion is likely to be relatively small and support the conclusions of Raskin et al. (1988), Horvath (1977), and Honts (1996). The present study was designed to test these two alternatives.

Method

Cases

Data collected from the U.S. Army Criminal Investigation Detachment Polygraph Division were used in this study. Criminal Investigation Detachment cases were selected because of the uniform procedures, high standards, and multiple levels of quality control implemented by that organization. Examiners in the Criminal Investigation Detachment have conducted polygraph exams throughout the United States and the world, wherever U.S. Army service members are assigned. About 20 field examiners and two quality control supervisors staff the Criminal Investigation Detachment Polygraph Division at any given time. All have field investigative experience, have at least a four-year college degree, are federally trained and certified, and meet continuing education requirements.

There are two important features of the U.S. Army Criminal Investigation Detachment investigative practices that merit comment. In that system, only those suspects who are the focus of the investigations are asked to submit to the polygraph examination. The polygraph is not used in a dragnet fashion. Also, all suspects are routinely confronted and interrogated by a Criminal Investigation Detachment criminal investigator a number of days before the polygraph examination is scheduled. Those who acknowledge the crime to the investigator are usually not polygraphed. It is these two pre-polygraph processes that might cause an increase of the proportion of guilty examinees in that polygraph population, and a decrease of the proportion of those predisposed to confess, more than in other systems with less examinee filtering.

From August 1996 through March 1998, U.S. Department of Defense Polygraph

Institute researchers reviewed all of the Criminal Investigation Detachment's polygraph cases for which ground truth confirmation could be found, beginning with cases conducted after January 1, 1995. The time period for the sampling was January 1, 1995, through February 3, 1997, when the last case meeting selection criteria was available to the researchers. During this period 3,349 polygraph examinations were conducted in criminal cases. Of these, 2,010 (60.0 percent) were calls of deception indicated, 884 (26.4 percent) were no deception indicated, and 455 (13.6%) were no opinion (inconclusive). There were 1,146 cases of examinee confessions, and no reports of false confessions.

Also reviewed were the investigative files for those polygraph cases that are maintained separately from the polygraph files and include details of all of the investigative and laboratory findings. Confirmation of the polygraph cases required at least one of the following: an unrecanted confession of the examinee, an unrecanted confession from someone who exculpated the examinee, evidence that the crime under investigation was never committed such as when missing property was discovered to have been innocently misplaced instead of stolen, forensic evidence such as urinalysis or surveillance tapes that substantiated the truth, or suspects led investigators to where they had hidden evidence or the stolen property. Eyewitness testimony, prosecutorial decisions, or judicial outcomes did not rise to the level of sufficient confirmation. Because, in the Criminal Investigation Detachment system, polygraph and other investigative measures were conducted concurrently rather than sequentially, discovery of evidence was somewhat more independent of the polygraph outcomes than in a system where the polygraph is used either very early or very late in the investigative process.

For consistency, polygraph examinations using a common testing format were selected for this study. The cases had to be single-issue examinations in which the U.S. Department of Defense Polygraph Institute zone comparison technique (U.S. Department of Defense Polygraph Institute 1992) was employed. Single-issue examinations are those in which a lie to one relevant question means the examinee lied to all relevant questions, or if truthful to one, was truthful to all. By U.S. Department of Defense Polygraph Institute standards, a minimum of three repetitions (charts) of the questions is required. If more than three charts were collected, only the first three complete charts were used in the study. By limiting the data in this fashion, the inconclusive rate for the samples was likely to have increased (Senter et al. submitted for publication), but it was seen as necessary to standardize the quantity of data from each case.

There were 704 examinations that met the criteria for polygraph format, scope (single issue), and a minimum number of charts. From that group, the authors obtained an in-depth sampling of 177 confirmed guilty cases where a confession was obtained from the examinee and 61 cases where the examinee did not confess, but other evidence established guilt. Of the 177 confessor cases, 28 had other supporting forensic evidence, and 149 were confession-confirmed only.

The complete review of the archived Criminal Investigation Detachment cases included a search for confirmed innocent cases meeting these criteria. For this study, an additional criterion was imposed on the innocent cases consistent with Raskin et al. (1988)- innocent cases had to come from multiple-examinee investigations in which the guilty party was discovered, or it was proven that the crime did not take place. Sixteen innocent cases were found to satisfy the multisuspect, scope, polygraph format, minimum chart, and ground truth criteria. Of these, five were theft cases in which the missing items were later discovered not to have been stolen, and the remaining cases were confirmed by the confession of someone other than the examinee. Examinee demographics for all cases meeting the selection criteria are found in Table 1.

Instrumentation

The Criminal Investigation Detachment polygraph program during this period used the Axciton computer polygraph (Axciton Systems, Incorporated, Houston, Texas) to record the traditional polygraph channels. There are two pneumographic sensors to register breathing, a standard blood pressure cuff for changes in blood volume, and finger electrodes for electrodermal activity. Data are digitized and available for offline analysis.

Table 1: *Examinee Demographics for the Criminal Investigation Detachment Sample*

	# of Cases	Males	Females	Suspects	Victims	Average Age
Confession Only	149	134	15	145	4	25.97 SD=7.22
Confession Plus Evidence	28	23	5	28	0	25.31 SD=4.45
Evidence Only	61	56	5	61	0	27.44 SD=5.70
Innocent	16	14	2	16	0	25.18 SD=6.83

Scoring Method

This study avoided the original examiners' scorings and decisions. They may have been prejudiced to some unknown extent by extra polygraphic sources of information such as case facts or the examinees' gestures and verbal behaviors (Iacono and Patrick 1987). The interest was in determining just how diagnostic the physiological data were when these extra polygraphic sources of information were excluded. A scoring method developed at the U.S. Department of Defense Polygraph Institute was chosen for this type of polygraph format, called the objective scoring system (Dutton 2000; Krapohl and McManus 1999). The objective scoring system uses physiological tracing features previously shown to be most diagnostic: respiration line length, electrodermal response amplitude, and blood volume amplitude (Kircher and Raskin 1988). Feature sizes for the relevant and comparison questions were converted into ratios where the measurement of each relevant question was divided by the measurement taken of the matched comparison question. The resultant ratios were compared to a chart of empirically developed thresholds for score assignment (Table 2). The scores were summed, and the totals were used for making a veracity decision.

Table 2: *Table for Conversion of Feature Ratios to Scores in the Objective Scoring System (Dutton 2000)*

Channel	Scoring Table						
RLL	0.00 - 0.79	0.80 - 0.89	0.90 - 0.96	0.97 - 1.03	1.04 - 1.10	1.11 - 1.25	1.26 – 999
Score	–3	–2	–1	0	1	2	3
EDR	999 - 2.45	2.44 - 1.61	1.60 - 1.21	1.20 - 0.93	0.92 - 0.68	0.67 - 0.44	0.43 - 0.00
Score	–6	–4	–2	0	2	4	6
BV	999 - 1.67	1.66 - 1.30	1.29 - 1.06	1.05 - 0.89	0.88 - 0.72	0.71 - 0.54	0.53 - 0.00
Score	–3	–2	–1	0	1	2	3

RLL = Respiration Line Length
EDR = Electrodermal Response
BV = Blood Volume

The objective scoring system scores for a three-chart polygraph examination have a potential range of -108 to +108. This system allows users to set their own cutting scores based on their tolerance for risk. The U.S. Department of Defense Polygraph Institute cutting scores of ±6 were used here: +6 or greater were categorized as no deception indicated, and -6 or lower were categorized as deception indicated. Scores between +/-6 were called inconclusive. These cutting scores produced decision accuracy at about 90 percent with the U.S. Department of Defense Polygraph Institute zone comparison technique (Krapohl and McManus 1999). The proportion of agreement between the trichotomous decisions of the objective scoring system and human blind scorers averaged 0.69 in that study.

Though the objective scoring system was designed to be performed manually in the field, the process was automated here to assure reliability. The three diagnostic features were measured automatically by a software package called Extract, version 3.0, developed for the U.S. government (Harris 1999). All had been conducted two years prior to the development of the objective scoring system; therefore, this scoring method had no influence on polygraph decisions by the original examiners or quality control personnel.

Results

Decision accuracies for each of the four groups are found in Table 3. Tests of proportions were conducted for each group to determine whether their accuracies exceeded chance expectancy of 0.50. In the first evaluation, decision errors and inconclusive decisions were both counted as errors. Each of the guilty groups produced detection rates above chance levels: confession only (z=5.49, p<.01), confession plus evidence (z=4.91, p<.01), and evidence only (z=4.23, p<.01). The detection rate for the 16 innocent cases was not greater than chance (z=1.50, p>.01). Tests of proportions that excluded inconclusive decisions found all four groups to have detection accuracy greater than chance: confession only (z=7.78, p<.01), confession plus evidence (z=4.91, p<.01), evidence only (z=5.63, p<.01), and innocent (z=3.33, p<.01).

The objective scoring system scores were evaluated for the three guilty groups, and a one-way ANOVA was calculated as a function of the group using scores as the dependent measure. The group effect was not significant (F[2, 235] = 0.58, p>.01). Figure 1 displays the mean scores, along with the standard error of measurement bars, for the three guilty groups and the one innocent group. The mean scores and standard deviations for the four groups are found in Table 4.

Table 3: *Decision Accuracy for Confession Only, Confession Plus Evidence, Evidence Only, and Innocent Cases Using the Objective Scoring System*

	Hit	**Miss**	**Inconclusive**	**Hit w/o Inconclusives**	**# of Cases**
Confession Only	72.5%	13.4%	14.1%	84.4%	149
Confession Plus Evidence	96.4%	3.6%	0.00%	96.4%	28
Evidence Only	77.1%	9.8%	13.1%	88.7%	61
Innocent	68.8%	6.3%	25.0%	91.7%	16

Figure 1: Mean Scores with the Standard Error of Measurement Bars for Confession Only, Confession Plus Evidence, Evidence Only, and Innocent Cases

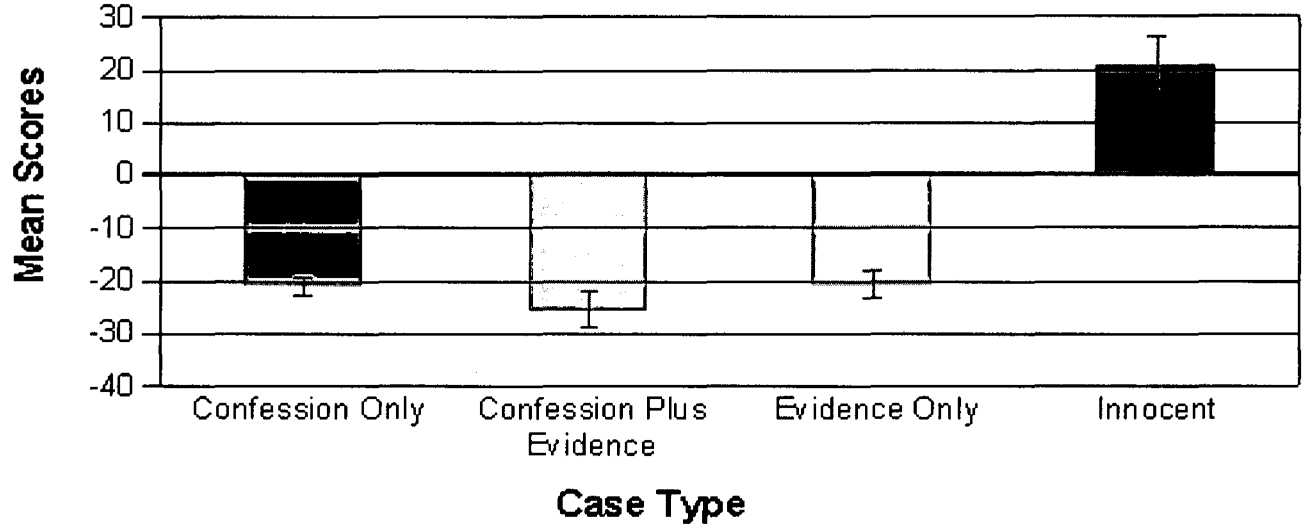


Table 4: *Means and Standard Deviations for Objective Scoring System Scores for Confession Only, Confession Plus Evidence, Evidence Only, and Innocent Cases*

	Mean	**SD**
Confession Only	-20.95	22.53
Confession Plus Evidence	-25.50	17.94
Evidence Only	-20.66	20.18
Innocent	20.56	20.77

Because there were no differences among the scores of the three guilty groups, those data were combined and a point-biserial correlation was conducted. Innocence was coded as 1 and guilt as 0. The correlation (*r*=0.43) was significant (*t* [252] = 7.65, *p*<.01).

Discussion

The present findings are consistent with the conclusions of Horvath (1977), Raskin et al. (1988), and Honts (1996). A liberal estimation with these datas' effect size, based on the one-way ANOVA, is quite small and negative due to the small value of the F ratio ω2 = -.015 (Keppel 1991). Taken in context with most of the other literature on the issue, this evidence should offer some reassurance to those who wish to undertake field research on the polygraph and the comparison question technique. However, the present conclusions are restricted to data that came from sources with practices similar to those of the U.S. Army Criminal Investigation Detachment.

The conclusions in the present data are at odds with the Patrick and Iacono (1991) findings. Both the present study and that of Patrick and Iacono (1991) used extensive field samples taken from law enforcement agencies, high-confirmation criteria, and independent analysis of the polygraph recordings, although there were significant methodological differences that limit what could be said about the discrepant findings. Patrick and Iacono relied on a semiobjective field-scoring system performed by human blind scorers, while the present study used an objective and automated method of scoring the data not available to Patrick and Iacono when their work was published. Also, the polygraph was not used as a last-ditch method of solving cases with the agency this study sampled, as Patrick and Iacono described the practice in their report. Therefore, it may have been easier to uncover ground truth for a larger proportion of cases in this study. The present study had the benefit of larger and possibly more homogenous samples, a more consistent polygraph testing protocol that had been monitored by quality control oversight, and digitized physiological data. And, while it should be noted that Patrick and Iacono's (1991) polygraph examiners used state-of-the-art examination procedures in the early 1980s when their data were collected in Vancouver, this study acknowledges that the practices of the more dispersed U.S. federal polygraph program in the 1990s are probably different.

The goal of this study was to determine whether there were differences in scores and decisions attributable to the confession criterion. Though none were found in this study, the confession criterion remains a potential source of contamination in undercontrolled studies. The present data demonstrate, however, that it is an overstatement to broadly assert that the confession criterion is a contaminant in a study. It is more defensible to state that the confession criterion is suspected when it leads to samples of cases with non-representative data, such as those with scores more extreme than the population as a whole. It should be relatively straightforward for researchers to collect and report such evidence as others have done so that skewed data can be recognized.

References

Ben-Shakhar, G., Lieblich, I., and Bar-Hillel, M. An evaluation of polygraphers' judgments: A review from a decision theoretic perspective, Journal of Applied Psychology (1982) 67(6):701-713.

Dutton, D. W. Guide for performing the objective scoring system, Polygraph (2000) 29(2):177-184.

Elaad, E., Ginton, A., and Ben-Shakhar, G. The effects of prior expectations and outcome knowledge on polygraph examiners' decisions, Journal of Behavioral Decision Making (1994) 7:279-292.

Faigman, D. L., Kaye, D. H., Saks, M. J., and Sanders, J. eds. Modern Scientific Evidence: The Law and Science of Expert Testimony. West, St. Paul, Minnesota, 1997.

Harris, J. C. Extract. Johns Hopkins University, Applied Physics Laboratory, 1999.

Honts, C. R. Criterion development and validity of the CQT in field application, Journal of General Psychology (1996) 123(4):309-324.

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

Horvath, F. The effect of selected variables on interpretation of polygraph records, Journal of Applied Psychology (1977) 62(2):127-136.

Iacono, W. G. Can we determine the accuracy of the polygraph tests? In: Advances in Psychophysiology. J. R. Jennings, P. K. Ackles, and M. G. H. Coles, eds. Jessica Kingsley, London, 1991, 4:202-208.

Iacono, W. G. The detection of deception. In: Handbook of Psychophysiology, 2nd ed., J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, eds. Cambridge University, New York, 2000.

Iacono, W. G. and Patrick, C. J. What psychologists should know about lie detection. In: Handbook of Forensic Psychology, I. B. Weiner and A. Hess, eds. Wiley, New York, 1987.

Keppel, G. Design and Analysis: A Researcher's Handbook. Prentice Hall, Englewood Cliffs, New Jersey, 1991.

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

Krapohl, D. J. and McManus, B. An objective method for manually scoring polygraph data, Polygraph (1999) 28(3):209-222.

Lykken, D. T. A Tremor in the Blood: Uses and Abuses of the Lie Detector. Plenum, New York, 1998.

Patrick, C. J. and Iacono, W. G. Validity of the control question polygraph test: The problem of sampling bias, Journal of Applied Psychology (1991) 76(2):229-238.

Raskin, D. C., Kircher, J. C., Honts, C. R., and Horowitz, S. W. A Study of the Validity of Polygraph Examinations in Criminal Investigation. Final report to the National Institute of Justice, Grant No. 85-IJ-CX-0040, 1988.

Senter, S. M., Dollins, A. B., and Krapohl, D. J. Comparison of Utah and DoDPI scoring accuracy: Equating veracity decision rule, chart rule, and number of data channels used. (submitted for publication).

U.S. Department of Defense Polygraph Institute. Zone Comparison Test. Fort McClellan, Alabama, 1992.

Exclusionary or nonexclusionary: A review of the evidence[1]

Donald J. Krapohl, Brett A. Stern and Andrew H. Ryan[2]

Abstract

In the field there has largely been a transition from nonexclusionary comparison questions to exclusionary questions over the past 40 years. The proponents of the exclusionary comparison question have persuasively argued that the inclusion of the relevant issue within the scope of the comparison question may cause the guilty examinee to consider those questions relevant, possibly resulting in false negative decisions. They argue that the clear distinction between the relevant and comparison question issues renders more accurate decisions. Conversely, the advocates for the nonexclusive comparison question have contended that much of the power of comparison questions resides in their ambiguity and expansiveness, and that narrowing the scope with exclusions necessarily makes them less, not more, effective. In the present paper we examine those arguments, present the available data, and assess the relative merits of both approaches.

"I have done that," says my memory. "I cannot have done that," says my pride and remains adamant – at last memory yields. Nietzsche.

In the beginning, there was the relevant-irrelevant technique. Though in common practice for decades, examiners recognized that it was encumbered with a false positive problem by virtue of the reliance on the mere presence or absence of physiological responses to relevant questions to assess deception. Examiners recognized early on that comparison questions were needed (Waller, 2001), and various forms emerged between the 1890s and the 1940s on a trial-and-error basis. The concept of the modern probable-lie comparison (PLC) question took hold in the late 1940s (Reid, 1947), and debates have since raged regarding what form is the most effective. Currently, the largest division among practitioners centers on whether PLCs should be devised so to avoid overlapping the relevant question (Horvath, 1988). The majority of polygraphers adhere to the method developed and espoused by Cleve Backster (Matte, 1996), namely, the exclusionary comparison question. The exclusionary PLC uses time-bars or other devices, with the aim of creating an unequivocal delineation between the relevant issue and the PLCs. An example of an exclusionary PLC for a 30-year-old suspected of a recent robbery might be: "Before the age of 29, did you ever steal anything?" By focusing the PLC on dishonest behavior in years before the current criminal act, exclusionary PLC proponents assert that the guilty examinee won't find the PLC to be a relevant question, a possibility that could dilute the effectiveness of the PLC. This confusion in the mind of the examinee, according to those following the Backster approach, could contribute to false negative decisions, the misclassification of a guilty examinee as truthful.

The opposing camp follow a method suggested and taught by John Reid. Reid used PLCs that were as broad as possible, which also typically encompassed the time of the crime being investigated. An example of these questions, called nonexclusionary (or inclusionary) PLCs, for the same robbery listed

[1] This paper is the second in a series of articles for a project entitled *Polygraph Myths*. The series will explore and try to verify assumptions about polygraphy held by practitioners and the public, through literature reviews and reanalysis of previous studies. For information on the project, contact the first author.

[2] The views expressed in this paper are those of the authors, and do not necessarily represent those of the Department of Defense or the US Government. The authors are with the US Department of Defense Polygraph Institute. Mr. Krapohl is Deputy Director, Mr. Stern is a Senior Instructor, and Dr. Ryan is Chief of Research Division. Request for reprints should be sent to the first author at: DoDPI, 7540 Pickens Ave., Ft. Jackson, SC 29207, or to krapohld@jackson-dpi.army.mil.

earlier could be: "Have you ever stolen anything in your whole life?" It is apparent that, if one were guilty of the robbery in question, one would also be lying to the nonexclusionary PLC. The use of nonexclusionary PLCs, sometimes called the "Reid method," is founded on the theory that broader and more general PLCs are more disconcerting, and thereby more effective in identifying nondeceptive examinees. They also believe them to be equally useful with guilty examinees, and scoff at the notion that examinees guilty of an offense would confuse a PLC that is generic and ambiguous with a relevant question that clearly specifies the particular offense by name and details.

Both the exclusionary and nonexclusionary PLCs were introduced many years before they were empirically tested. The first direct comparison of the two approaches was reported by Podlesny and Raskin (1978). In a mock-crime analog study in which they evaluated each polygraph data channel individually, they concluded that the exclusionary PLC was superior to the nonexclusionary PLC. The finding rested on advantages the exclusionary PLC offered in mean skin conductance recovery half-time, mean skin conductance response recovery half-time width, and negative skin potential response amplitudes. Interestingly, decision accuracy for the exclusive and nonexclusive PLCs were only different by two cases, which was not statistically significant. There could be several reasons for finding that physiological phenomena were affected by the type of PLC, but polygraph decision accuracy wasn't. One might simply be the modest sample sizes in the Podlesny and Raskin study, which could have lacked the sensitivity to detect a difference in decision accuracy if it really existed. A compelling case could also be based on the observation that the physiological measures that did show differences between exclusionary and nonexclusionary PLCs are not the same ones used in manual scoring. The features normally used in manual scoring did not show any significant differences between the exclusionary and nonexclusionary PLCs. Therefore, Podlesny and Raskin's careful research might have uncovered a statistically significant effect, but the type of PLC could very well have little or no practical effect in field polygraph examinations.

Horvath took up the topic of PLC type again in 1988. Like Podlesny and Raskin (1978), Horvath conducted a mock-crime analog study. However, Horvath's research question focused on the PLC's influence on scores and decisions. His findings were that nonexclusionary PLCs produced scores more in the correct direction than did exclusionary PLCs and rendered fewer false positives. However, overall decision accuracies for the two types of PLCs were not significantly different. Horvath's data also contradicted a common assertion that nonexclusive PLCs are less effective with guilty examinees because they overlap the relevant issue. The nonexclusive PLCs actually generated more deceptive scores than did the exclusive PLCs with guilty examinees. At first blush, these findings appeared to be in conflict with Podlesny and Raskin (1978) whose data favored the exclusionary PLCs. The disagreement might not be as large as it appears, however. Though both Horvath (1988) and Podlesny and Raskin (1978) determined that some fundamental components of individual scores may be affected, both converged on the practical finding that there were no significant differences in decision accuracy between the two types of PLCs.

Subsequent research extended the work of Horvath (1988). As part of his Masters thesis at Michigan State University, Palmatier (1991) conducted a laboratory study with 120 subjects in which the exclusionary and nonexclusionary PLCs were independent variables. In addition, Palmatier compared the accuracy of the Zone Comparison Technique (ZCT) and the Modified General Question Technique (MGQT), the two most prevalent forms of probable-lie comparison question tests. Palmatier found that decision accuracy was not affected by the technique. Neither the MGQT nor the ZCT had statistically different accuracies. However, unlike previous studies, Palmatier found the type of PLC had a strong influence on accuracy, errors, and inconclusive rates. In all three categories, the nonexclusionary PLC was superior to the exclusionary PLC. This effect was most pronounced for the innocent subjects, who enjoyed a much lower error rate with the nonexclusionary PLC. In the discussion of his findings, Palmatier (1991) acknowledged that the rationale for the time-bar approach seemed plausible, but that his data ran firmly against it.

The three studies agreed that there was no advantage in decision accuracy for the exclusionary PLC, however, to determine whether the findings would generalize out of the laboratory, field data were needed. In 1999, Israeli researcher Tuvya Amsel reported his findings for 230 field cases for which he had ground truth. The study used three different examiners who did not share the same polygraph training, and who used the 3-position scoring system. The 3-position scoring system is an adaptation of the 7-position scoring system (Capps & Ansley, 1992; Krapohl, 1998; Van Herk, 1990), but with a restricted range that may serve to reduce the level of subjectivity in score assignments. Amsel determined that the use of nonexclusionary PLCs generated higher mean absolute scores than did the use of exclusionary PLCs for both of two relevant questions used in a form of ZCT. In other words, scores for the guilty examinees were more negative, and scores for the innocent were more positive, when the nonexclusionary PLCs were used instead of the exclusionary PLCs. Table 1 summarizes the findings of the four different studies discussed here.

Table 1.

Study	Study Type	Sample Sizes Exclusive	Sample Sizes Non-Exclusive	PDD Techniques	Better Accuracy	Details
Podlesny & Raskin (1978)	Lab	20	20	ZCT	Equal	Exclusive PLCs had more discriminative responses in some physiological measures, but no effect on decision accuracy
Horvath (1988)	Lab	20	20	MGQT	Equal	Detection accuracies were not significantly different. Nonexclusive PLCs produced fewer errors with both innocent and guilty
Palmatier (1991)	Lab	60	60	MGQT & ZCT	Nonexclusive	Nonexclusive PLCs had higher accuracy and produced fewer false positives
Amsel (1999)	Field	87	143	ZCT	Nonexclusive	Nonexclusive PLCs rendered stronger average scores in the correct direction for both innocent and guilty

Each of the four cited studies took different approaches to the problem, which incurred different methodological strengths and liabilities. There were two polygraph techniques, blind scorers with dissimilar training, two types of scoring systems, and very different samples of subjects. Had these researchers arrived at conflicting conclusions regarding the relative decision accuracy of the exclusionary and nonexclusionary PLC, it would not have been a great surprise. However, none found an increase in decision accuracy attributable to the exclusionary PLC, converging instead on the finding that exclusionary PLCs are, at least, no better than the nonexclusionary PLCs. We were unable to locate any relevant studies, published or unpublished, that concluded otherwise.

Prolific polygraph researcher and writer Dr. Charles Honts summarized the evidence:

> ... (T)he idea of probable-lie control has evolved over the years. The original Reid controls were things like "have you ever told a lie" or "have you ever done anything that was dishonest or illegal." That evolved into a time-bracketed control, which was something that Cleve Backster introduced in the early '60's. And the idea was to separate the probable-lie controls from the relevants in time. So there were time-bars put on them -- before the age of 35 or before 1994 -- some way to separate that from the issue. Backster's concern being that if the control and relevants overlap, that may be confusing for the subject. It actually turns out that is probably wrong....Reid controls seem to work just as well as the Backster time-barred controls, although most polygraph examiners do the time-bar. It really does look like from science that is not necessary. (Honts, 1996).

Trade-offs

Exclusionary PLCs

The scientific issues notwithstanding, there may be some advantages to exclusionary PLCs, at least for those who give testimony on polygraph cases. First, the logic for exclusionary PLCs remains appealing. And, if their ready acceptance by field practitioners is any guide, they may receive a more favorable hearing from judges and juries. Second, the number of validity studies using exclusionary PLCs is far larger than that of nonexclusionary PLCs. In laying the foundation for evidence, the use of exclusionary PLCs in the case would help avoid the additional step of testifying to the evidence on how the type of PLC really doesn't matter.

All benefits come with costs, of course. For the exclusionary PLCs, they may be the ease in which they are recognized by an examinee who wants to target them as part of a countermeasure scheme. A cursory review of the countermeasure advice offered on the Internet or in books shows that examinees are encouraged to look for markers of exclusionary PLCs, such as the time bars. Avoiding these unmistakable signs increases the difficulty for the potential counter-measurer to identify the target. Though nonexclusionary PLCs can also be spotted by the sophisticated examinee, nonexclusionary PLCs do not carry the buzzwords that make them quite as obvious.

Finally, exclusionary PLCs are no prophylactic against false negatives. While they provide a line of demarcation between the relevant and the comparison issues, it is questionable whether the PLC is truly applicable to a given subject, and if applicable, whether that applicability is of such importance that it overwhelms the relevant issue. Polygraph practitioners may find such a phenomenon occurring in serial offenders - particularly with burglars and rapists.

Nonexclusionary PLCs

One of the principles of PLC question construction is that the question should be as general as possible. Advocates of the CQT agree that a focused PLC would be less effective than a broad PLC. If this principle is true, modifiers added to a PLC may serve to narrow them, reducing their salience to the innocent examinee for whom these questions were developed. Horvath (1988), Palmatier (1991), and Amsel (1999) all found such an effect in the polygraph scores. Exclusionary PLCs reduced the differential arousals between relevant questions and PLCs as compared to the nonexclusionary PLCs. It is not unreasonable to attribute part of the decrement in effectiveness to the narrowing of the scope of the PLCs with time bars. Therefore, nonexclusionary PLCs may perform better in practice, despite the accepted wisdom of the exclusionary PLC.

One disadvantage of the nonexclusionary PLC is that they are easier for the less trained polygraph examiner to mishandle. A poorly constructed or improperly introduced nonexclusionary PLC may, indeed, turn into a relevant question, increasing the probability of a decision error. As an example, in the polygraph examination of a suspect in a burglary, an ill-conceived nonexclusionary PLC could be: "Have you ever stolen anything from a building?" Though that question satisfies the requirement of being general, it may be too close to the relevant issue to be the proper choice of PLC, illustrating the need for competent selection of PLCs. The use of time bars is one way of making a definite demarcation for the examinee between relevant and comparison questions, and because they are easier for the novice polygraphers to administer properly, they present an important benefit.

Conclusion

Exclusionary PLCs are the standard for most polygraph schools, many state and local governments, and at this writing, almost all US Federal polygraph programs. An overwhelming majority of research investigating polygraph validity has used the exclusionary PLC. From the field evidence and laboratory findings, it may safely be concluded that exclusionary PLCs are very effective. The evidence abruptly halts, however, before one can assert that the exclusionary PLCs are more effective than nonexclusionary PLCs. By consensus, the research finds this claim false.

Though it is a scientific cliché to suggest more research on an issue at the end of a paper, it does not seem warranted in this case. With a 0-4 record, the notional superiority of the exclusionary PLC is probably a settled matter. This does not discount its usefulness, of course. Polygraphers and agencies can rely instead on the nonscientific considerations cited earlier to continue a preference for this type of comparison question.

References

Amsel, T.T. (1999). Exclusive or Nonexclusive comparison questions: A comparative field study. *Polygraph, 28*(4), 273-283.

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

Horvath, F. (1988). The utility of control questions and the effects of two control question types in field polygraph techniques. *Journal of Police Science and Administration, 16*(3), 198-209.

Honts, C.R. (1996). Testimony in *United States v. Gilliard,* US District Court, Southern District of Georgia, Augusta Division, Case No. CRI9E-19. Full text found at http://truth.boisestate.edu/polygraph/honts01.html

Krapohl, D.J. (1998). A comparison of 3- and 7- position scoring scales with laboratory data. *Polygraph, 27*(3), 210-218.

Matte, J.A. (1996). *Forensic Psychophysiology Using the Polygraph: Scientific Truth Verification – Lie Detection.* J.A.M. Publications: Williamsville, New York.

Palmatier, J.J. (1991). *Analysis of Two Variations of Control Question Polygraph Testing Utilizing Exclusive and Nonexclusive Controls.* Unpublished masters degree thesis, Michigan State University, East Lansing, Michigan.

Podlesny, J.A., & Raskin, D.C. (1978). Effectiveness of techniques and physiological measures in detection of deception. *Psychophysiology, 15*(4), 344-359

Reid, J.E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law and Criminology, 37*, 542-547.

Van Herk, M. (1990). Numerical evaluation: Seven point scale +/-6 and possible alternatives: A discussion. *The Newsletter of the Canadian Association of Police Polygraphists, 7*(3), 28-47.

Waller, J.F. (2001). A concise history of the comparison question. *Polygraph, 30*(3), 192-195.

Modified General Question Test Decision Rule Exploration

Stuart M. Senter

Abstract

Previous studies by Senter, Dollins, and Krapohl (2000), Senter (2001a, 2001b), and Wygant (2003) showed that substantial increases in accuracy could be gained within the Zone Comparison Test format by using different novel decision rule combinations. Such modifications have shown to produce large increases in the number of correct decisions with nondeceptive participants, both with laboratory and field data with the Zone Comparison Test. The current work focused on boosting accuracy through the modification of the decision rule process with the Modified General Question Test. In this effort, 205 field cases across four different data sets were used. Results of the current study showed that the use of 'total score' approaches with the MGQT could produce more balanced accuracy results across deceptive and nondeceptive cases, relative to conventional 'spot total' approaches. Overall, accuracy rates were higher for deceptive cases than for nondeceptive cases. Results are discussed in the context of base rates and polygraph testing objectives.

The field of the Psychophysiological Detection of Deception (PDD) is extremely complex, involving the coordination of several processes and procedures in order to produce diagnostic decisions. An overarching goal that should be inherent in the field is the continual search and exploration of the optimal manner in which to implement the various components of the PDD process. Through the diligent and thorough scrutiny of these components, we will continue to elevate the performance and credibility of the field. This project explored a single piece of these procedures, with a very specific PDD format: decision rules with the Modified General Question Test (MGQT).

MGQT Format

The MGQT (Department of Defense Polygraph Institute [DoDPI], 2002) is a probable-lie comparison (PLC) testing format. From a decision-making standpoint, PLC formats are comprised of relevant questions and comparison questions. Relevant questions probe direct involvement or secondary participation of a specific incident (i.e., Did you steal any of that money?). Comparison questions query the past behaviors of the examinee that are categorically similar to the relevant issue, typically separated by a time bar or some other qualifier (i.e., Prior to this year, did you ever steal anything?). Relevant questions are generally paired with one or two comparison questions, and the relative response magnitudes of these questions are compared using respiratory, electrodermal, and cardiovascular measures. If larger responses are produced following the relevant question in a given pairing, negative values are assigned. If larger responses are produced following the comparison question, positive values are assigned. These assigned values are summed across physiological channels (respiratory, electrodermal, and cardiovascular) and question repetitions and are used to produce decisions of deceptive (large negative totals), nondeceptive (large positive totals), or inconclusive (totals close to zero).

The MGQT format is distinct from the other common PLC format, the Zone

[1]Address correspondence to Stuart Senter, Department of Defense Polygraph Institute, Fort Jackson, SC 29207 or Senters@jackson-dpi.army.mil. This research was funded by the Department of Defense Polygraph Institute, Fort Jackson, South Carolina, as project DoDPI02-P-0008. The views expressed in this report are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Comparison Test (ZCT) (DoDPI, 2002), in that the scope of the MGQT is broad, containing multiple aspects of a given crime or incident. The MGQT typically contains a single primary relevant question, which asks directly whether or not the examinee performed the crime, and one or more secondary relevant questions which probe for knowledge, participation, or other involvement in the crime (i.e., Do you know who stole that money?). The MGQT is also less structured than the ZCT, which follows a strict sequence of questions and standard number of relevant and comparison questions. Instead, the MGQT may include from 2-4 relevant questions and 2-5 comparison questions, contingent upon the scope of the incident in question. The MGQT also excludes a symptomatic or outside issue question (e.g., Are you concerned that I will ask you a question that we have not reviewed?) that is typically included to encompass any extraneous areas of concern in the narrower-scope ZCT.

Finally, of particular relevance to the present study, the MGQT is different from the ZCT in that individual question pair totals are the sole determination of deceptive or nondeceptive decisions, as opposed to the sum of all question totals. The conventional threshold for producing a nondeceptive decision for the MGQT is a minimum of +3 for each relevant/comparison pair, and a value of -3 or lower for any single relevant/comparison pair for producing a deceptive decision.

Previous MGQT Research

Previous research on the MGQT has generally shown a high accuracy rate, with some variation, typically with higher accuracy rates for deceptive cases than for nondeceptive cases. Previous studies have provided decision performance using both conventional decision thresholds for the MGQT and collapsed question totals, typical of the ZCT. With this latter approach, all assigned values are summed (even across spot totals) and decision thresholds are applied (typically -6/+6). Table 1 shows the results of studies using conventional MGQT thresholds (spot totals) and those using total scores. Although conventional wisdom would suggest that using total scores is inappropriate given the broad scope of relevant questions contained within the MGQT format, both decision approaches perform reasonably well. The weighted mean percent correct excluding inconclusive decisions for the spot totals approach was 84.8%, with an 18.6% inconclusive percentage. These percentages were comparable for the total scores approach at 80.1% and 15.3%, respectively.

The present study will explore the effectiveness of various spot total cutoffs, total cutoffs, and combinations of these two approaches using four data sets. Previous work by Senter (2001b) and Wygant (2003) on ZCT decision approaches showed increases in decision accuracy and reductions in inconclusive decisions with decision rules that differed from conventional rules. A similar approach was taken in the present study, in order to determine whether a more effective decision rule approach exists that may be different from the conventional rule employed with the MGQT.

Method

Polygraph Cases

A total of 205 verified criminal specific field cases (161 deceptive, 44 nondeceptive) were used in the study. Twenty-six polygraph cases (21 deceptive, five nondeceptive) were collected from the United States Army Criminal Investigation Division Command (USACIDC) and 47 cases (44 deceptive, three nondeceptive) collected during the years of 1998 and 1999 were provided by the Bureau of Alcohol, Tobacco, and Firearms (ATF). Thirty-two cases (16 deceptive, 16 nondeceptive) were used from a study by Krapohl and Norris (2000). Finally, 100 cases (80 deceptive, 20 nondeceptive) were used from a study by Blackwell (1998). These cases were used for accuracy assessment and cross validation as a function of the various decision rules. Specific crimes and demographic information were not available for most cases, and thus were not explored as variables in the present study.

Table 1
Cutoffs and Accuracy Rates for Prior MGQT Studies

			Percent Accuracy Excluding Inconclusives				
Decision Type and Study	N	Cutoffs	Dec.	Ndec.	Weighted Total	Averaged* Total	Inc.
Spot Totals							
Blackwell (1998)	100	+3/-3	96.7	31.9	88.5	64.3	7.0
Crowe et al. (1998)	30	+3/-3	85.7	100.0	92.9	92.9	53.3
Honts & Barland (1990)	88	+3/-3	90.5	62.1	78.9	76.3	19.3
Jones & Salter (1989)	9	+3/-3	100.0	100.0	100.0	100.0	11.1
Krapohl & Norris (2000)	32	+3/-3	100.0	13.8	65.3	56.9	25.0
Podlesny & Truslow (1993)	96	+2/-3	93.8	64.3	88.6	79.1	17.7
Total Scores							
Horvath (1988)	40	+5/-5	unkn.	unkn.	79.5	79.5	2.5
Palmatier (1991)	60	+6/-6	unkn.	unkn.	69.4	69.4	18.3
Podlesny & Truslow (1993)	96	+6/-6	84.8	94.7	87.2	89.8	18.8

Note. N = total number of participants, Cutoffs = for 'Spot Totals', this is the total required for each question to produce a nondeceptive decision and total required for any individual question to produce a deceptive decision, respectively, Dec. = Deceptive, Ndec. = Nondeceptive, Inc. = Inconclusives. unkn. = Specific values unknown. * = this represents the average of deceptive and nondeceptive accuracy rates.

Data Reduction

The polygraph charts from the USACIDC and the ATF were all scored by the original examiners. The charts from Krapohl and Norris (2000) and Blackwell (1998) were each blind scored by three evaluators. Decisions by each set of three evaluators from these latter two studies were averaged following the application of each decision rule. For each relevant/comparison pair and for each channel (respiratory, electrodermal, and cardiovascular), a value was assigned according to the differential responses produced by the two questions (see Swinford, 1999, for a description of the scoring criteria used to produce the data sets in this study). A value of +1 was assigned if the response to the comparison question was greater than the response to the relevant question. A value of -1 was assigned if the response to the relevant question was greater than the response to the comparison question. A zero was assigned to the question pair if the responses to the two questions were not different. The number of relevant/comparison pairs varied from two to four. Each sequence of questions was repeated three times. For the purposes of the present study, only the totals for each relevant/comparison pair (spot totals) and the sum of all assigned scores (total scores) were taken into consideration.

Decision Rules

Table 2 displays cutoffs and descriptions of the decision rules that use only spot totals to produce decisions. Rules 1-6 are modifications of the conventional MGQT scoring rule which requires a value of +3 for each question pair in order to produce a decision of no deception indicated (NDI), and a value of -3 for any question pair to produce a decision of deception indicated (DI). Rule 1 starts with +1/-1 cutoffs, and these increment to maximum of +6/-6 at Rule 6. Rules 7-12 are balanced (BAL) with respect to producing DI versus NDI decisions. Both require the same decision thresholds for each question pair. For example, BAL1 requires a +1 or higher value in each spot to produce a decision of NDI and requires -1 or lower in each spot to produce a decision of DI. Rules 13-16 explore the impact of using asymmetric (ASYM) cutoffs for DI and NDI thresholds. Rules 17 and 18 represent the typical MGQT decision rule with the exception that the threshold for positive values necessary to produce an NDI decision is adjusted (ADJ).

Table 2
Cutoffs and Decision Rule Descriptions for Spot Cutoff Approaches

Rule Code	Rule Name	Spot Cutoffs	Total Cutoffs	Description
1	MGQT1	+1/-1	None	+1 required in each spot for NDI, -1 in any spot for DI
2	MGQT2	+2/-2	None	+2 required in each spot for NDI, -2 in any spot for DI
3	MGQT3	+3/-3	None	+3 required in each spot for NDI, -3 in any spot for DI
4	MGQT4	+4/-4	None	+4 required in each spot for NDI, -4 in any spot for DI
5	MGQT5	+5/-5	None	+5 required in each spot for NDI, -5 in any spot for DI
6	MGQT6	+6/-6	None	+6 required in each spot for NDI, -6 in any spot for DI
7	BAL1	+1/-1	None	+1 required in each spot for NDI, -1 required in each spot for DI
8	BAL2	+2/-2	None	+2 required in each spot for NDI, -2 required in each spot for DI
9	BAL3	+3/-3	None	+3 required in each spot for NDI, -3 required in each spot for DI
10	BAL4	+4/-4	None	+4 required in each spot for NDI, -4 required in each spot for DI
11	BAL5	+5/-5	None	+5 required in each spot for NDI, -5 required in each spot for DI
12	BAL6	+6/-6	None	+6 required in each spot for NDI, -6 required in each spot for DI
13	ASYM1	+1/-5	None	+1 required in each spot for NDI, -5 required in each spot for DI
14	ASYM2	+2/-4	None	+1 required in each spot for NDI, -5 required in each spot for DI
15	ASYM3	+5/-1	None	+1 required in each spot for NDI, -5 required in each spot for DI
16	ASYM4	+4/-2	None	+1 required in each spot for NDI, -5 required in each spot for DI
17	ADJ1	+1/-3	None	+1 required in each spot for NDI, -3 required in any spot for DI
18	ADJ2	+2/-3	None	+2 required in each spot for NDI, -3 required in any spot for DI

Note. MGQT = Modified General Question Test, BAL = Balanced, ASYM = Asymmetric, ADJ = Adjusted, NDI = No Deception Indicated, DI = Deception Indicated.

Table 3
Cutoffs and Decision Rule Descriptions for Spot Cutoff Approaches

Rule Code	Rule Name	Spot Cutoffs	Total Cutoffs	Description
19	ZCT1	+1/-3	+6/-6	+6 total and +1 required in each spot for NDI, -6 total or -3 in any spot for DI
20	ZCT2	None	+6/-6	+6 total required for NDI, -6 total for DI
21	ZCT3	+1/-3	+6/-6	application of ZCT1 rule, followed by application of ZCT2 rule if inconclusive
22	ZCT4	+1/-3	+6/-6	application of ZCT2 rule, followed by application of ZCT1 rule if inconclusive

Note. ZCT = Zone Comparison Test, NDI = No Deception Indicated, DI = Deception Indicated.

Table 3 displays decision rules that incorporate total cutoffs in order to produce decisions. Given the precedent of using total scores with the MGQT (bottom panel of Table 1), these four approaches which rely on total scores were included. Rules 19 and 20 explore the impact of conventional ZCT decision rules, using the DoDPI spot score rule and total cutoffs, respectively. The ZCT1 rule requires a value of +1 or higher in each question spot, and a summed value of +6 or higher to produce a decision of NDI. To produce a decision of DI, a value of -3 or lower in any spot or a total score of -6 or lower is required. The ZCT2 rule is the only rule considered in the current set that looks at total scores exclusively, with no reliance on spot totals. Finally, rules 21 and 22 represent two-stage rules (Senter, 2001b), which alternate use of rules 19 and 20 in sequence. The second stage of rules 21 and 22 is only used in the event of an inconclusive decision produced in the first stage.

Procedure

Each decision rule was applied to the assigned scores of the USACIDC cases using spreadsheet software, producing decisions of DI, NDI, and inconclusive. The performance of each rule was tabulated for each individual study. In addition, collapsing across the four data sets, a signal detection analysis (Macmillan & Creelman, 1991) was conducted in order to obtain a measure of sensitivity and bias for each decision rule. As a final analysis, ground truth was coded as +1 and -1 for nondeceptive cases and deceptive cases, respectively. These values were correlated with decisions produced, with nondeceptive, deceptive, and inconclusive decisions coded as +1, -1, and 0, respectively. This 'decision efficiency index' (DEI) was calculated using Spearman rank order correlation coefficients (Cohen & Cohen, 1983).

In the present study, an important consideration was performance for each decision rule, not only within a particular study, but across all studies. Thus, an elimination process was instituted whereby for each data set, each decision rule was required to achieve minimum thresholds of performance in order to be considered an effective rule. First, decision rules had to achieve a minimum total (deceptive and nondeceptive combined) accuracy rate of 70%, excluding inconclusive decisions. Second, a given decision rule could have no more than 30% inconclusive decisions. This second elimination component was put in place to eliminate those rules which may produce 100% accuracy, but resolve only a small proportion of cases.

Results

Decision Performance

Tables 4-7 display the decision performances for each rule meeting the exclusion criteria across the four data sets. Rule inclusion was

Table 4
Decision Rule Performance for CID Data Set (N = 26)

	Percent Accuracy Excluding Inconclusives				
Decision Rule	Deceptive (n=21)	Nondeceptive (n=5)	Weighted Total	Averaged Total	Inconclusive
ZCT1	100.0	75.0	95.7	87.5	11.5
ZCT2	94.1	75.0	90.5	84.6	19.2
ZCT3	95.0	75.0	91.7	85.0	7.7
ZCT4	95.0	75.0	91.7	85.0	7.7
MGQT1	100.0	75.0	95.8	87.5	7.7
MGQT2	100.0	75.0	95.8	87.5	7.7
MGQT3	100.0	66.7	95.5	83.4	15.4
ASYM2	100.0	75.0	95.0	87.5	23.1
ASYM3	100.0	50.0	95.5	75.0	15.4
ASYM4	100.0	50.0	95.5	75.0	15.4
ADJ1	100.0	75.0	95.7	87.5	11.5
ADJ2	100.0	75.0	95.7	87.5	11.5

stable across each data set, with the exception of the Krapohl and Norris (2000) study, where only the ZCT2, ZCT3, and ZCT4 rules met the exclusion criteria. It should be noted that this was the only study with a high base rate of nondeceptive cases (50%). The other studies were extremely unbalanced in this regard (16% average base rate of nondeceptive cases). For each individual study, accuracy rates for deceptive cases excluding inconclusive decisions exceeded 90 percent for each decision rule that met the cutoff criteria,

Table 5
Decision Rule Performance for ATF Data Set (N = 47)

	Percent Accuracy Excluding Inconclusives				
Decision Rule	Deceptive (n=44)	Nondeceptive (n=3)	Weighted Total	Averaged Total	Inconclusive
ZCT1	95.3	100.0	95.7	97.7	2.1
ZCT2	90.6	100.0	91.4	95.3	25.5
ZCT3	93.2	100.0	93.6	96.6	0.0
ZCT4	93.2	100.0	93.6	96.6	0.0
MGQT1	95.5	100.0	95.7	97.8	0.0
MGQT2	97.6	100.0	97.8	98.8	4.3
MGQT3	97.6	100.0	97.8	98.8	4.3
MGQT4	100.0	100.0	100.0	100.0	25.5
ASYM2	97.1	100.0	97.4	98.6	19.2
ASYM3	100.0	0.0	100.0	50.0	10.6
ASYM4	100.0	100.0	100.0	100.0	10.6
ADJ1	95.3	100.0	95.7	97.8	2.1
ADJ2	97.6	100.0	97.8	98.8	4.3

Table 6
Decision Rule Performance for Krapohl & Norris Data Set (N = 32)

	Percent Accuracy Excluding Inconclusives				
Decision Rule	Deceptive (n=16)	Nondeceptive (n=16)	Weighted Total	Averaged Total	Inconclusive
ZCT2	85.3	65.7	75.4	75.5	28.1
ZCT3	93.6	45.5	70.3	69.6	5.2
ZCT4	89.3	52.3	71.4	70.8	5.2

again, with the exception of the Krapohl and Norris data. However, for nondeceptive cases, only the ZCT2 and ZCT4 rules exceeded 50 percent for each individual study, and only the ZCT2 rule exceeded 60 percent for each study. This accuracy imbalance between deceptive and nondeceptive cases is reflected in Table 8 which shows the percentage of correct, erroneous, and inconclusive decisions as function of ground truth for those decision rules that met the exclusion criteria for at least one data set. Across all rules, the percentage of correct decisions for deceptive cases was very high, but very low for nondeceptive cases.

Signal Detection Analysis

The decisions produced by each of the 13 decision rules listed in Table 8 were subjected to a signal detection analysis to discern the effectiveness of each rule beyond simple decision performance. In order to produce binary responses necessary for the analysis, inconclusive decisions were treated as missing data, with deceptive and nondeceptive cases and decisions coded as 0 and 1, respectively. Critical measures reported for this analysis are *d'* and Log(*b*) values produced by each decision rule. The parameter *d'* is a sensitivity index, of an individual decision rule in this

Table 7
Decision Rule Performance for Blackwell (1998) Data Set (N = 100)

	Percent Accuracy Excluding Inconclusives				
Decision Rule	Deceptive (n=80)	Nondeceptive (n=20)	Weighted Total	Averaged Total	Inconclusive
ZCT1	100.0	30.2	89.1	65.1	8.3
ZCT2	98.5	72.5	94.1	85.5	20.3
ZCT3	98.7	47.4	88.7	73.1	2.7
ZCT4	98.7	50.9	89.4	74.8	2.7
MGQT1	100.0	25.5	86.9	62.8	3.7
MGQT2	100.0	25.5	87.5	62.8	6.3
MGQT3	100.0	21.1	88.9	60.6	10.0
MGQT4	100.0	10.7	89.5	55.4	20.7
ASYM2	100.0	32.4	89.9	66.2	17.7
ASYM3	100.0	5.0	86.3	52.5	7.3
ASYM4	100.0	7.9	87.1	54.0	9.3
ADJ1	100.0	30.2	89.1	65.1	8.3
ADJ2	100.0	28.6	89.1	64.3	8.7

Table 8
Decision Percentages for All Four Data Sets (N = 205) as a Function of Decision Rule

	Deceptive (n = 161)			Nondeceptive (n = 44)			Total Accuracy Excluding Inconclusives	
Decision Rule	Corr.	Err.	Inc.	Corr.	Err.	Inc.	Weighted	Averaged
ZCT1	94.4	1.7	3.9	31.8	43.2	25.0	88.5	70.3
ZCT2	74.5	4.1	21.3	53.0	19.7	27.3	90.3	83.8
ZCT3	94.4	3.7	1.9	49.2	43.2	7.6	87.4	74.8
ZCT4	94.0	4.1	1.9	53.0	39.4	7.6	87.4	76.6
MGQT1	97.1	1.7	1.2	31.8	55.3	12.9	86.3	67.4
MGQT2	95.5	0.6	3.9	28.8	50.8	20.5	87.7	67.8
MGQT3	94.4	0.6	5.0	21.2	43.2	35.6	89.0	66.2
MGQT4	81.8	0.0	18.2	9.1	37.1	53.8	89.3	59.9
ASYM2	81.8	0.6	17.6	28.8	37.1	34.1	89.3	71.5
ASYM3	97.1	0.0	2.9	5.3	55.3	39.4	86.7	54.4
ASYM4	95.5	0.0	4.6	9.1	50.8	40.2	87.6	57.6
ADJ1	94.4	1.7	3.9	31.8	43.2	25.0	88.5	70.3
ADJ2	94.4	0.6	5.0	28.8	43.2	28.0	89.2	69.7

Note. Corr. = Correct, Err. = Erroneous, Inc. = Inconclusive.

instance, basically representing the separation between the distribution of deceptive cases and the distribution of nondeceptive cases that a decision rule produces. The larger the value of *d'*, the greater the separation between the two distributions. Log(*b*) is an index of response bias, with greater deviations from zero indicating larger tendencies to respond in a biased manner (i.e., mostly deceptive decisions, even with nondeceptive cases).

Table 9 shows the *d'* and Log(*b*) values produced for each decision rule. The ASYM2 rule produced the greatest *d'* value, followed by the ADJ2, ZCT2, MGQT2, and MGQT3 rules, once again, indicating greater separation between the distribution of deceptive and nondeceptive cases. However, on

Table 9
Signal Detection Results: d' and Log(b) Values as a Function of Decision Rule for the Combined Data Sets

Decision Rule	*d'*	Log(*b*)
ZCT1	1.88	-2.12
ZCT2	2.18	-1.11
ZCT3	1.90	-1.57
ZCT4	1.89	-1.43
MGQT1	1.74	-2.10
MGQT2	2.16	-3.04
MGQT3	2.02	-2.97
MGQT4	1.83	-3.22
ASYM2	2.30	-2.95
ASYM3	1.28	-2.67
ASYM4	1.70	-3.19
ADJ1	1.88	-2.12
ADJ2	2.25	-3.05

Table 10
Decision Efficiency Index as a Function of Decision Rule Combining Across the Four Data Sets

Decision Rule	Decision Efficiency Index
ZCT1	.571
ZCT2	.530
ZCT3	.589
ZCT4	.593
MGQT1	.532
MGQT2	.548
MGQT3	.572
MGQT4	.432
ASYM2	.459
ASYM3	.513
ASYM4	.542
ADJ1	.571
ADJ2	.578

the whole, great separation was accompanied by a great deal of bias, with the ASYM2, ADJ2, MGQT2, and MGQT3 rules also showing the greatest Log(*b*) deviations (excepting the ASYM4 and MGQT4 rules, which produced lower *d'* values and the greatest Log(*b*) deviations). The distinct exception to this trend was the ZCT2 rule which produced the lowest Log(*b*) deviation from zero, followed by the ZCT4 and ZCT3 rules.

Decision Efficiency Index

As a final measure of the effectiveness of each decision rule, the DEI was calculated for the decision rules in Table 8. This measure correlates decisions (coded as 1, 0, and -1, for nondeceptive, inconclusive, and deceptive decisions, respectively) and ground truth (coded as 1 and -1 for nondeceptive and deceptive cases, respectively). Spearman correlations are calculated and the correlation coefficients represent efficiency indices, whereby higher values indicate higher efficiency. Correct decisions increase this index, while incorrect and inconclusive decisions decrease it. Table 10 shows the results of this analysis. All DEI values were statistically significant. The highest DEI's were produced by the ZCT4 and the ZCT3 rules. The highest DEIs for 'non-ZCT' rules were produced by the ADJ1, the ADJ2, and the MGQT3 rules.

Discussion

Limitations

Before presenting a discussion of results, it is important to note a number of limitations inherent in this study. First, only field cases were used in the present study, many of which were confirmed by confession only. As Iacono (1991) argued, this approach can lead to overestimations of polygraph accuracy, due to the systematic exclusion of false positive and false negative errors. A more representative approach would be to use polygraph cases that have been verified through means independent of the polygraph decision. A second limitation of the present study is sample size. Polygraph is an extremely robust procedure, but is subject to examinee and examiner variability, in addition to variations in the nature of each specific polygraph case (another limitation of using field data exclusively). Thus, confidence in polygraph results can be derived only through the use of extremely large sample sizes.

A third potential limitation was the overall base rates in the present study, which were extremely unequal on the side of deception (78.5% of cases). The potential problem with this is that if a given decision approach were to simply make a deceptive call on each and every case, an accuracy rate of 78.5% would be achieved. Given the imbalanced results of the ZCT2 approach (+6/-6 total score cutoffs only), the MGQT data in the present study certainly appears to be biased toward producing deceptive decisions, and thus performed extremely well with the present data set, with the exception of the Krapohl and Norris (2000) data. The implication is that with a balanced base rate (i.e., 50% deceptive, 50% nondeceptive), the MGQT would not produce such a high total accuracy rate.

However, there are two additional issues to address in the context of this limitation. First, in many criminal specific cases, the polygraph is only used in those instances where the list of potential suspects has been narrowed considerably, or in cases where the polygraph is the 'final option' in a case. Thus, a high base rate of deception among criminal specific examinees is likely to exist in this context. Second, the tendency of the MGQT to produce deceptive decisions can be useful, even in the context of low deceptive base rate examinee populations, if implemented properly. For example, the MGQT may be useful as an initial test to explore possible areas of concern (in the context of a pre-employment examination), or to explore different components of a specific criminal incident. This test could then be followed with a more specific diagnostic test such as the ZCT. Thus, the MGQT may be useful as an initial stage of testing, with the purpose of directing the content of a more specific test. The exploration of such an approach is beyond the scope of the present study, but would be an important area to research in future studies.

A fourth limitation is the mixture of blind scoring data with original examiner data. The Blackwell (1998) and Krapohl and Norris (2000) data were each produced by three independent blind evaluators, while the USACIDC and ATF data were produced by the original polygraph examiners. As Iacono (1991) indicates, it is likely difficult for original examiners to exclude non-polygraph data, in some manner, into the decision process, thus illustrating the importance of using blind evaluation to gain an estimate of the validity of the polygraph process.. The potential impact of this variable can be discerned most clearly by the substantially lower accuracy rates for nondeceptive cases produced by the 'blind evaluator' data (Tables 6 and 7) compared to the 'original examiner' data (Tables 4 and 5). However, statistical tests for such differences are fairly meaningless due to the small proportion of nondeceptive cases throughout all data sets (excepting the Krapohl and Norris data), and the small proportion in the 'original examiner' data sets in particular (8 out of 73 cases). In addition, as indicated previously, the base rates in the Krapohl and Norris study were very different from the other three data sets, and may also have impacted the accuracy rates. Clearly, a more consistent approach would have been to use blind evaluations for all data sets.

A fifth limitation of the present study had to do with the inconsistency in the structure of the MGQT cases used across the various studies. The MGQT cases in the Blackwell (1998) and ATF data sets used from 2-4 relevant questions, while the cases used in the Krapohl and Norris (2000) and USACIDC data sets used 4 relevant questions exclusively. The impact of this variation is unknown.

'Optimal' Decision Rules

This study explored a variety of decision rules across four individual studies to explore whether an optimal rule or set of rules would arise. Decision performance in the present study shared many similarities with that of the previous studies shown in Table 1. First, use of the MGQT results in a large discrepancy in decision performance between deceptive cases and nondeceptive cases. This discrepancy can be partially explained by the imbalance in decision thresholds for deceptive and nondeceptive classifications. Producing deceptive decisions using traditional MGQT decision rules is much easier than producing nondeceptive decisions. The former requires only a single question spot to reach a value of -3 or lower and the latter requires each question spot to attain a value of +3 or higher. However, this explanation is not entirely satisfactory, because even the completely balanced ZCT2 rule showed a large discrepancy in accuracy between deceptive and nondeceptive cases. This suggests that the very structure of the MGQT is biased toward the production of deceptive decisions.

In the present study, there was no perfect decision rule, nor one that produced the best decision accuracy across all contexts. As Tables 4-8 indicate, there are a number of decision rules that performed well with deceptive cases, only to fall short with nondeceptive cases. However, those that performed moderately well with nondeceptive cases (i.e., ZCT2, ZCT3, and ZCT4) were slightly less accurate with deceptive cases than other rules. Further, the combined results of the signal detection analysis and the implementation of the DEI did not produce a clear choice. While the ZCT2 rule produced the best combination of sensitivity and bias, other approaches performed better with the DEI, most likely due to the relatively high percentage of inconclusive decisions produced by the ZCT2 rule.

The results produced by 'total score' rules (i.e., ZCT1-4) in the present study are consistent with the results of previous studies that have applied such rules to the MGQT. Overall, this class of decision rules performed well, with particular improvement over the 'MGQT' class of rules with nondeceptive cases, accompanied by a slight decrement in relative performance with deceptive cases. The multiple issue nature of the MGQT raises fundamental questions about using total scores to produce decisions, provided the possibility that there may be large imbalances in spot totals. In spite of this criticism, the more balanced performance produced by the 'Total Score' rules, as evidenced by decision results, signal detection analysis, and the DEI analysis suggests that these rules may be worthy of consideration in certain contexts.

Two issues to consider in the classification of one rule over another with respect to performance are the goal of the test and, once again, the potential base rate of the examinee population. As suggested above, the MGQT as an approach appears to be highly biased towards producing deceptive decisions. Thus, its use as a stand-alone diagnostic instrument may be problematic with respect to the production of false positive errors. However, if, as suggested above, the goal of the test is to identify areas for further investigation, and not to produce a final decision regarding the truthfulness of the examinee, the MGQT may be well-suited for such a role. In such cases, false positives will have the opportunity to be rectified in a follow-up examination, such as the ZCT. As a preliminary test, the MGQT is likely superior to the ZCT in this capacity, due to an increased probability of identifying one of a set of issues for further investigation, whereby the narrow scope of the ZCT has an increased probability of missing such issues.

The estimated base rate of an examinee population clearly has a great impact on the effectiveness of a given decision rule. In the present study, the base rate of deception was extremely high. Decision rules with a tendency or bias toward producing deceptive decisions were extremely effective with this sample. However, in the context of populations with lower base rates of deception, overall accuracy rates will decrease for such decision rules. In such contexts, decision rules with an extreme bias toward will reduce the diagnostic value of the polygraph test, if the test is used as a stand-alone procedure. In such cases where only a single examination is possible, it would certainly seem preferable to implement the use of the more balanced 'ZCT-type' rules, or to avoid the use of the MGQT altogether.

In conclusion, we have at our disposal a wide array of decision rules that may be implemented. The decision of what decision rule to use, in addition to which test format to conduct should be driven by the estimated base rate of the examinee population, with the ultimate goal of maximizing the diagnostic value of the polygraph procedure. By doing so, even if it means the implementation of less than traditional decision rules, the validity of the process can be elevated.

References

Blackwell, N. J. (1998). PolyScore 3.3 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations. *Polygraph, 28*(2) 149-175.

Cohen J. & Cohen P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Crowe, M. J., Chimarys, M., & Schwartz, J. R. (1998, August). *The GQT polygraph test: Scoring and validity.* Poster presented at the 96th annual convention of the American Psychological Association.

Department of Defense Polygraph Institute (2002). *Modified General Question Test.* Instructional pamphlet for Psychophysiological Detection of Deception Course.

Department of Defense Polygraph Institute (2002). *Zone Comparison Test.* Instructional pamphlet for Psychophysiological Detection of Deception Course.

Honts, C. R. & Barland, G. H. (1990). *A laboratory study of the validity of the MGQT: An executive summary.* Fort McClellan, Alabama: Department of Defense Polygraph Institute.

Horvath, F. S. (1988). The utility of control questions and the effects of two control question types in field polygraph techniques. *Journal of Police Science and Administration, 16*(3), 198-209.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? *Advances in Psychophysiology, 4*, 201-207.

Jones, H. E. & Salter, S. (1989). Polygraph accuracy: An analog study. *Polygraph, 18*(2), 69-74.

Krapohl, D. J. & Norris, W. F. (2000). An exploratory study of traditional and objective scoring systems with MGQT field cases. *Polygraph, 29*(2),185-194.

Macmillan, N. A. & Creelman, C. D. (1991). Detection theory: A user's guide. New York, NY: Cambridge University Press.

Palmatier, J. J. (1991). *Analysis of two variations of control question polygraph testing utilizing exclusive and nonexclusive controls.* Unpublished master's thesis, Michigan State University.

Podlesny, J. A. & Truslow, C. M. (1993). Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology, 78*(5), 788-797.

Senter, S. M. (2001a, May). *Improving accuracy and utility of specific issue polygraph: New possibilities.* Talk given at the annual Forensic Sciences and Crime Scene Technology Conference, Washington, DC.

Senter, S. M. (2001b, July). *New approaches for increasing polygraph accuracy.* Talk given at the annual meeting of the American Polygraph Association, Indianapolis, IN.

Senter, S. M., Dollins, A. B., & Krapohl, D. J. (2000). *Comparison of Utah and DoDPI scoring accuracy: Equating veracity decision rule, chart rule, and number of data channels used.* (Report No. DoDPI00-R-0001). Fort Jackson, SC: Department of Defense Polygraph Institute.

Swinford, J. (1999). Manually scoring polygraph charts using the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph, 28*(1), 10-27.

Wygant, J. R. (2003). Title. Scoring cutoffs – picking the best. *Polygraph, 32*(2), 86-96.

Instructions to Authors

Scope

The journal *Polygraph* publishes articles about the psychophysiological detection of deception, and related areas. Authors are invited to submit manuscripts of original research, literature reviews, legal briefs, theoretical papers, instructional pieces, case histories, book reviews, short reports, and similar works. Special topics will be considered on an individual basis. A minimum standard for acceptance is that the paper be of general interest to practitioners, instructors and researchers of polygraphy. From time to time there will be a call for papers on specific topics.

Manuscript Submission

Manuscripts should be in English, and submitted, along with a cover letter, to Editor, American Polygraph Association, PO Box 10342, Ft. Jackson, South Carolina 29207 (USA). The cover letter should include a telephone number, return address, and e-mail address. Authors should also state clearly in the cover letter if they wish to submit their manuscript to a formal peer-review. The preferred method of manuscript submission is as an email attachment (MS Word, WordPerfect, or PDF format) with the cover letter included in the body of the email. Send to the Editor at:

pollinad@att.net

Authors without Internet access may also submit manuscripts on computer disk along with 5 paper copies to the editorial address above. As a condition for publication, authors shall be required to sign a statement that all text, figures, or other content in the submitted manuscript is correctly cited, and that the work, all or in part, is not under consideration for publication elsewhere.

Manuscript Organization and Style

All manuscripts must be complete, balanced, and accurate. All authors should follow guidelines in the *Publication Manual of the American Psychological Association* (4th edition). The manual can be found in most public and university libraries, and can be ordered from: American Psychological Association Publications, 1200 17th Street, N.W., Washington, DC 20036, USA. Authors are responsible for assuring their work includes correct citations. Consistent with the ethical standards of the discipline, the American Polygraph Association considers quotation of another's work without proper citation a grievous offense. The standard for nomenclature shall be the *Terminology Reference for the Science of Psychophysiological Detection of Deception* included in this volume. Legal case citations should follow the *West* system.

Manuscript Review

A single Associate Editor will handle papers, and the author may, at the discretion of the Associate Editor, communicate directly with him or her. For all submissions, every effort will be made to provide the author a review within 12 weeks of receipt of manuscript. Articles submitted for publication are evaluated according to several criteria including significance of the contribution to the polygraph field, clarity, accuracy, and consistency.

Copy-editing

The Editor reserves the right to copy-edit manuscripts. All changes will be coordinated with the principal author.

Copyright

Authors submitting a paper to the American Polygraph Association (APA) do so with the understanding that the copyright for the paper will be assigned to the American Polygraph Association if the paper is accepted for publication. The APA, however, will not put any limitation on the personal freedom of the author(s) to use material contained in the paper in other works, and request for republication will be approved, if the senior author concurs.

Professional Copies

The senior author will receive ten (10) copies of the journal issue in which the article appears.