Controversies

Psychophysiological Detection of Deception

Charles R. Honts

The psychophysiological detection of deception (PDD; also known as polygraphy or lie detection) has been, and remains, an important application of psychology in the real world. These psychophysiological tests involve the recording of autonomic nervous system indices (e.g., respiratory, electrodermal, cardiovascular, and vasomotor activity) while the subject is asked a series of questions about which his or her credibility is being evaluated. PDD tests are used extensively in the criminal justice system of the United States, and most police departments have a polygraph examiner to support their investigative divisions. Under some circumstances, the results of PDD tests are admissible as evidence in courts of law, and confessions obtained from suspects subsequent to PDD examinations are frequently offered as evidence.[1] Furthermore, despite the fact that many of the private-sector industrial uses of PDD tests were outlawed in 1988, these tests remain an important part of the American workplace.[2] The use of PDD examinations for preemployment screening of police officers is widespread at all levels of government, and PDD examinations are an important part of the process for obtaining and maintaining a security clearance in the U.S. government.

Charles R. Honts is an Associate Professor of Psychology at the University of North Dakota. Address correspondence to Charles R. Honts, Psychology Department, P.O. Box 8380, University of North Dakota, Grand Forks, ND 58202-8380.

All the applications of PDD examinations are controversial. Since the early 1970s, a polemic debate has persisted in the scientific literature concerning basic and applied questions about using psychophysiological measures for the purpose of detecting deception. This controversy has been particularly heated in discussions of the control question test (CQT), the most commonly applied PDD technique in the criminal justice system. Scientists are split both over the use of laboratory studies as a basis for estimating validity and over what criteria to use in determining what an adequate field study is.[3] The ins and outs of this extensive debate are beyond this review. However, three new areas of research have recently provided evidence that bears upon this debate, and I address them here. First, I consider a series of studies and analyses that provide data concerning the validity of PDD tests used in the national security system. Second, I address research on a new PDD technique, the directed-lie control test, that has been found to have high criterion validity, but avoids most of the conceptual criticisms of the CQT. Finally, I address research on the application of computer-based statistical decision making in the psychophysiological detection of deception and consider the use of bootstrapping as a decision-making technique.

PDD TESTS USED FOR SECURITY SCREENING

In terms of numbers of examinations conducted, the industrial applications of PDD tests are by far more important than the criminal justice applications. Moreover, current industrial uses have particular importance because PDD tests determine who becomes a law enforcement officer, who has access to our government's most closely guarded secrets, and who has access to and control over our most dangerous weapons. Despite the importance placed on the outcomes of such tests, before the late 1980s, virtually no scientific studies on the use of PDD tests in the workplace were conducted. In 1987, the U.S. Department of Defense (DOD) began a series of studies on the accuracy of the various PDD techniques used for national security screening. Prior to those studies, both the scientific critics and the proponents of PDD had predicted that PDD tests used for national security screening would produce large numbers of false positive errors.[4] However, the DOD studies were to show that the critics and proponents were both wrong.

In 1987, DOD began a large study of the validity of four of the PDD techniques used by the U.S. government for national security screening.[5] The 207 subjects of this experiment were all federal employees; 74% of them had access to classified materials at some point in their careers. Polygraph examiners from Military Intelligence (Army INSCOM), the Air Force Office of Special Investigations, the National Security Agency, and the Central Intelligence Agency conducted the polygraph examinations using their standard screening techniques. About half of the volunteer subjects in this study were innocent, and about half were made "guilty" either by having them enact a mock crime of espionage or by providing them with knowledge about acts of mock espionage.

The mock-crimes subjects participated in were very complex, and some required multiple acts over periods of time as long as 6 weeks.

Real-world spy handlers controlled the subjects as they would operatives in the real world. The subjects went to clandestine meetings with foreign-sounding agents in local bars. Code words were exchanged to initiate contact. Subjects entered government facilities and copied realistic mock-classified documents and then made dead drops in the local community. They then received money from their handlers for their espionage activities. Subjects in the knowledge condition received the same experiences and information as the mock-espionage subjects, but never had the opportunity to carry out the acts.

The mock-espionage and knowledge scenarios provided the kind of experiences the national security screening program was designed to uncover. The subjects were told that under no circumstances were they to reveal their involvement in the experiment to anyone other than the experimenters, and they were told to resist interrogation by the polygraph examiners as long as possible. In addition, they were told that if they revealed that they had been involved in real espionage, they might be arrested and subjected to criminal prosecution.

The examiners in this study were good at obtaining information from the subjects: Twenty percent of the subjects made admissions about real-world security violations (not part of the scenarios), and some of those admissions were serious enough that they had to be classified. However, as is shown in Table 1, the original examiners were not very good at detecting deception. In fact, their judgments of guilt or innocence accounted for only 10% of the true variation in guilt or innocence. Independent evaluations based only on the physiological data (i.e., with no knowledge of the case facts or the subjects' verbal and nonverbal behavior during the pretest portion of the original examination) were performed by quality control officers from each of the parent agencies. Those independent evaluations accounted for only 4% of the criterion variance, and only one agency produced better-than-chance results.[6] The most striking finding of this study was not that the accuracy of national security screening tests was low, but that most errors were false negative errors. That is, a large number of guilty people passed their PDD examinations.[7]

These are not the only data that support the conclusion that national security PDD examinations are not accurate and produce large numbers of false negative errors. DOD has reported data on the outcomes of 67,049 PDD examinations given in the Counterintelligence Scope Polygraph Program. I have previously conducted conditional probability analyses on these data, and an update of those analyses is shown in Table 2. This analysis used 20% as the base rate of deception, a rate obtained from DOD's own estimates.[5] Data from the field can be seen to closely converge with data from the laboratory, with the screening tests showing low validity and a preponderance of false negative outcomes. This analysis suggests that in the field, 95.4% of the deceptive individuals are classified incorrectly, while the error rate with innocent individuals cannot be larger than 0.11%.

These results are in striking contrast to the general trend of scientific commentary over the years and to the results of conditional probability analyses based on the outcomes of research on forensic applications of PDD. How is it possible that all the scientific commentary and the analyses were wrong? Numerous interviews I have conducted with government examiners suggest an answer. The examiners reported knowing that there are few true targets (actual spies). They are told by their supervisors (and by Congress) not to falsely accuse innocent people, and they are put under pressure to maintain low rates of "deceptive" and "inconclusive" outcomes. Under these circumstances, it does not seem surprising that they have altered their testing and decision-making procedures to ensure that they call very few subjects deceptive.

These results strongly suggest that, as currently applied, polygraph tests used for security screening are not effective. However, it might be possible to use PDD in screening as part of a series of successive hurdles. In such an application, the polygraph would be used to narrow the field of applicants. Then, other tests or investigations would be conducted to distinguish true positive from false positive outcomes. Additional research in this area is clearly needed.

Table 1. *Outcomes for the original examiners in the Department of Defense's study of national security screening polygraph tests*

Condition	Examination outcome		
	Truthful	Inconclusive	Deceptive
Innocent	105	4	7
Guilty (mock crime or knowledge)	55	8	28

Note. Data are from Barland, Honts, and Barger.[5]

THE DIRECTED-LIE CONTROL TEST

The most common polygraph technique in the field is the CQT. With the CQT, inferences about a

Table 2. *Results of a conditional probability analysis of data from the Department of Defense's Counterintelligence Scope Polygraph Program*

Condition	Examination outcome		
	Truthful	Deceptive	Total
Innocent	53,293	**60**	53,353
Guilty	13,066	**630**	13,696
Total	**66,359**	**690**	**67,049**

Note. Numbers shown in boldface are empirical; other numbers were derived.

subject's veracity are made on the basis of differential responding to relevant and control questions. Relevant questions are designed to address the issue under investigation, and they should be clear and unambiguous (e.g., "Did you take the money from the safe?"). Control questions are designed to be broad and ambiguous, and the examiner maneuvers the subject into answering them "no" without making admissions (e.g., "Before 1993, did you ever do anything that was dishonest, illegal, or immoral?"). The term "control question" is a misnomer. In the CQT, control questions are designed to actively elicit strong physiological responses from innocent subjects, rather than to serve as controls in the scientific sense.

Subjects are led to believe that physiological responses to the control questions will be viewed as negatively as physiological responses to the relevant questions. The rationale of the CQT assumes that individuals attempting deception on the central issues of the examination will respond physiologically to the relevant questions. It is further assumed that although innocent individuals will realize that the relevant questions are important, they will respond with greater physiological responses to the control questions. The latter assumption is based on the reasoning that the innocent know they did not commit the crime addressed by relevant questions, but they are either lying or at least uncertain in their responses to the control questions. Guilty subjects are not expected to respond as strongly to the control questions as to the relevant questions because they are lying to the relevant questions and the relevant questions are much more important.

In the evaluation of a CQT, if a subject's physiological responses occur primarily to relevant questions, the subject is judged to be deceptive. If a subject's physiological responses occur primarily to the control questions, the subject is considered truthful. Lack of response, or undifferentiated responses to relevant and control questions, results in an inconclusive outcome.

The rationale of the CQT has been widely criticized as unreasonable, unworkable, unethical, and not supportable empirically, whereas proponents argue that the rationale is reasonable, is easy to implement, and has strong empirical support.[3] Central to the criticisms of the CQT seem to be two concerns. First is a concern that it is not reasonable to think that polygraph examiners can develop control questions that will compete successfully with the relevant questions when the subject is innocent. Thus, the CQT is seen as doomed to produce a large number of false positive outcomes. The second major concern about the CQT is that developing control questions for each subject during the interview introduces a strong subjective element into the examination. To some critics, this subjective element and idiographic approach raise serious questions about the basic nature of the CQT as a test. Those critics have argued that the CQT is so subjective that it can neither be standardized nor even specified adequately for research purposes. Furthermore, this subjective element also raises the specter of possible manipulation of the examination outcome by the examiner.

One possible response to this debate is to examine other techniques that may avoid the criticisms leveled at the CQT. One potential replacement for the CQT is the directed-lie control test (DLCT). The rationale of the DLCT is similar to that of the CQT except that the comparison question, the one expected to elicit response from the innocent, is a known lie. For example, the examiner may ask, "Have you ever told a lie, even one time in your life?" The subject initially answers "yes," but is then directed to answer "no" during the examination. In the DLCT, truthful and deceptive subjects are expected to respond differentially to the relevant and directed-lie questions.

The directed-lie control questions are prepared in the following manner. A subject is told that it is important for comparison purposes that he or she answer some of the questions on the test deceptively. The examiner also tells the subject that it is critical that he or she respond appropriately when lying. However, the nature of appropriate responding is not defined for the subject. Finally, the subject is told that if he or she does not react appropriately to the directed-lie questions, the examination will be inconclusive and will have to be repeated at another time. In this case, differential reactivity is expected because the innocent subject's attention has been focused on the directed-lie questions by the examiner's instructions and by concern over responding appropriately. The DLCT is evaluated in the same manner as the CQT.

The DLCT avoids the two principle objections to the CQT. First, no attempt to balance the content of the

comparison questions with the relevant questions is needed. The power of a directed-lie question is imparted by the instructions, not by the specific content of the question. The second criticism is also addressed because the test is standardized. Every subject can receive the same instructions and the same directed-lie questions. Thus, concerns about the adequacy of comparison questions are eliminated. The DLCT is sufficiently standardized that it may even be possible to administer it with a machine.

Initial research results with the DLCT are very promising. In a laboratory mock-crime experiment, my colleagues and I contrasted two kinds of directed lies, those that involved personally relevant information (e.g., "Have you ever told a lie even one time in your life?") and trivial directed lies that did not involve the subject personally (e.g., "Does 2 + 2 = 4?"), with the CQT and an older technique known as the relevant-irrelevant test. The results of that experiment are summarized in Table 3.[8] To assess the discriminative classification power of the various techniques, we calculated a detection efficiency coefficient (*r*) for each. The square of the correlation coefficient can be used to give an index of the percentage of variance accounted for in the guilt criterion. The DLCT using the personal directed lies produced the greatest absolute discrimination between the innocent and guilty subjects. Although decisions with the personal directed lie accounted for 16% more of the variance in the guilt criterion than decisions with the CQT, this difference did not reach statistical significance. This lack of effects was probably due to a lack of power in the design.

Similar, but stronger, results were reported in a field study that I conducted with David Raskin. In 1984, we began including a directed-lie control question in every polygraph test we conducted in our private practices. These tests were conducted for a variety of clients and covered a range of crimes from child sexual abuse to homicide. In early 1987, we took an exhaustive sample of all of our cases that had been confirmed innocent or guilty by confession or by incontrovertible physical evidence developed after the polygraph examination. The polygraph charts were then subjected to a blind independent analysis.[9] The principal result of that study is illustrated in Figure 1. The inclusion of the directed-lie control question produced a significant effect on the numerical scores. Mean numerical scores for innocent subjects were shifted in the positive direction, while the scores of the guilty subjects were little affected by the inclusion of the directed-lie control question. This effect was reflected in the blind decisions. When the directed lie was used, there were no false positive outcomes with this sample of cases. When the directed lie was not used, 15% of the outcomes were false positive errors.

Data from both the field and the laboratory indicate that the directed-lie control question is at least as effective as traditional control questions, and the results from the field suggest that the use of the directed lie may reduce the number of false positive errors produced. These results, combined with the clear conceptual and psychometric advantages of the DLCT, make a strong case for its use in the field.

Table 3. *Outcomes of a study of various kinds of control questions*

Control technique and guilt condition	Examination outcome			Detection efficiency *r*
	Truthful	Inconclusive	Deceptive	
Control				.56
Innocent	12	1	2	
Guilty	3	4	8	
Personal directed lie				.69
Innocent	13	0	2	
Guilty	2	3	11	
Trivial directed lie				.50
Innocent	10	3	2	
Guilty	3	4	8	
Relevant-irrelevant				.38
Innocent	3	1	11	
Guilty	0	0	15	

Note. Data are from Horowitz, Raskin, Honts, and Kircher.[8]

STATISTICAL DECISION MAKING IN THE DETECTION OF DECEPTION

One long-standing and interesting finding of the decision sciences has been that statistical decision making often outperforms expert human decision makers. Following in that tradition, in the late 1970s, Raskin and his graduate students (primarily John Kircher) began work on developing and validating a computer-based statistical decision procedure for use with the CQT. It was felt that such a procedure would be more powerful, objective, and reliable than were the evaluations made by polygraph examiners. The hope of removing subjective biases from the decision process was particularly important. Starting with laboratory data, Kircher and Raskin used discriminant analysis to develop optimal linear equations to discriminate innocent and guilty subjects. The discriminant scores that resulted from those equations were then en-

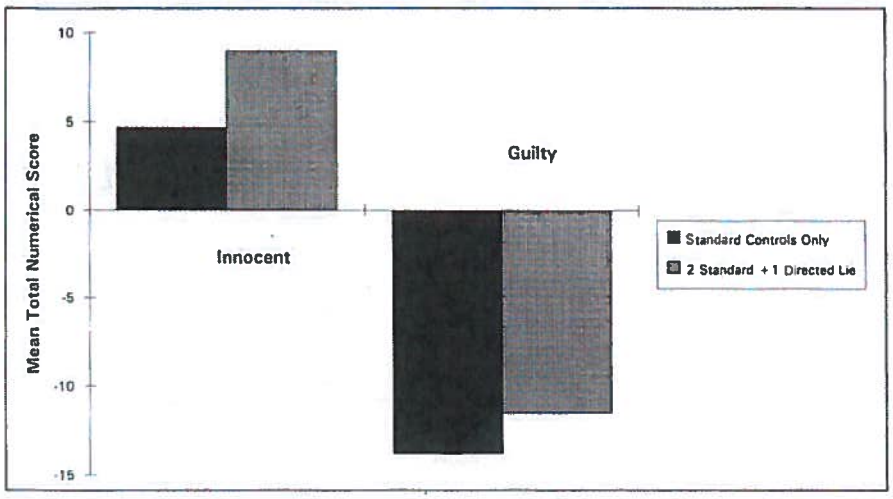

Fig. 1. Mean total numerical scores, for innocent and guilty subjects, from a field study of polygraph examinations evaluated both with and without a directed-lie control question. Each examination contained three relevant questions, two standard control questions, and one directed-lie control question. Two independent numerical evaluations of physiological responses were performed. In the first evaluation, the directed-lie control question was ignored, and each relevant question was compared with the nearest standard control question. In the second evaluation, the directed-lie control question was included in the numerical evaluation for purposes of scoring the nearest relevant question. Scores were assigned on a 7-point scale that ranged from −3 to +3; negative scores indicate greater physiological response to the relevant questions, and positive scores indicate greater physiological response to the control questions. The magnitude of the difference in response between the question types was reflected in the magnitude of the scores. Data are from Honts and Raskin.[9]

tered into Bayes' Theorem to calculate a probability of truthfulness for each subject. The resulting discriminant model has been cross-validated with laboratory and field data and has been compared with expert human evaluators. In every case, the discriminant model has evaluated the physiological data as well as, or better than, the best human experts.[10]

Although powerful, the discriminant analysis approach has limitations. Discriminant analysis makes a number of mathematical and distributional assumptions that may be hard to meet in field data. What is more important, the generalizability of discriminant models is limited by the size and representativeness of the specific data sets upon which they are built. Current work with discriminant analysis and PDD has been criticized on these grounds.

Mary Devitt and I have recently explored the possibility of using a bootstrapping approach to decision making as an alternative to discriminant analysis. Bootstrapping is a resampling approach that uses massive computation to build estimates of parameters from sample data. Bootstrapping seemed an attractive alternative to discriminant analysis because it makes no assumptions of the data and does not rely on a specific sample of data as the basis for model building.

In our study, we examined the relative validity of bootstrapping, discriminant analysis, and expert human evaluators with a sample of 100 innocent and 100 guilty subjects from previous mock-crime experiments.[11] The results of this study are summarized in Table 4. Across the three methods of evaluation, the results were statistically equivalent. However, it should be noted that the human evaluators in this study were all highly experienced psychophysiologists. Their results are likely to overestimate the average accuracy in the field, as the training and sophistication of field examiners have been generally criticized.[1]

These results may be of great interest to people studying the detection of deception as well as to people interested in decision making in general. Even with performance equivalent to that of discriminant analysis, the bootstrap approach to decision making offers the following advantages:

- Issues regarding generalizing from a model built on a particular standardization sample to other populations are moot. The results of the bootstrap are very likely to be generalizable because they are based only on the data from the subject at hand.
- The bootstrap makes no mathematical assumptions of the data.

Table 4. *Outcomes of a study of three approaches to decision making in the detection of deception*

Analytic technique and guilt condition	Examination outcome			Detection efficiency *r*
	Truthful	Inconclusive	Deceptive	
Human numerical				.65
Innocent	70	24	6	
Guilty	9	38	53	
Discriminant analysis				.67
Innocent	68	25	7	
Guilty	12	19	69	
Bootstrap				.71
Innocent	71	17	12	
Guilty	8	13	79	

Note. Data are from Honts and Devitt.[11]

- The bootstrap is a general approach and should be adaptable to a variety of tests for detection of deception and to other statistical decision-making situations.

SUMMARY AND CONCLUSIONS

Current trends in research on the psychophysiological detection of deception suggest that progress is being made in moving from being strictly a practice and clinical art to being a science based on standardization, psychometrics, and statistics. New techniques such as the DLCT and the use of statistical decision making are removing the subjective clinical element and are replacing it with sound, reliable, and valid methods. The standing of the PDD tests in courts of law may increase as this process advances. Current trends in the law suggest that this change has already begun.[1]

However, current research on the use of polygraph tests in national security points out that many serious problems still exist in the field. The surprising finding that most of the errors made in the screening context are false negative errors suggests that a great deal of additional research is going to be needed before polygraph tests can play a useful role in that setting. Until that work is done, it would seem prudent to limit rather than expand the use of the polygraph as a screening device.

Clearly, much additional work needs to be done in this area. It is my hope that more psychologists will become involved in this important research. Moreover, as the research base expands, I hope the polemics will decrease and the rate of progress will increase.

Acknowledgments—I would like to thank Sandra Scarr, John Kircher, Mary Devitt, and two anonymous reviewers for their suggestions and comments during the development of this manuscript.

Notes

1. For a review of the legal status of polygraph tests in the United States, see C.R. Honts and M.V. Perry, Polygraph admissibility: Changes and challenges, *Law and Human Behavior, 16*, 357–379 (1992). The probability of increased offers and admissibility of polygraph test results would seem to be greatly enhanced by a recent U.S. Supreme Court decision that supplanted the long-controlling and conservative *Frye* standard for the admissibility of scientific evidence with the more liberal relevance standards of the Federal Rules of Evidence; see Daubert v. Merril Dow Pharmaceuticals, Inc., 113 S.Ct. 2786 (1993).

2. Most private-sector industrial uses were outlawed by the *Employee Polygraph Protection Act of 1988*, Public Law 100-347, 29 U.S.C. §2001 (1988). For a review of the industrial uses of the polygraph test, see C.R. Honts, The emperor's new clothes: Applications of polygraph tests in the American workplace, *Forensic Reports, 4*, 91–116 (1991). PDD examinations are also used in other countries; for a review, see G.H. Barland, The polygraph test in the US and elsewhere, in *The Polygraph Test: Lies, Truth, and Science*, A. Gale, Ed. (Sage, Beverly Hills, CA, 1988).

3. The anti-PDD side of this debate is typified by D.T. Lykken, *A Tremor in the Blood: Uses and Abuses of the Lie Detector* (McGraw-Hill, New York, 1981), and G. Ben-Shakhar and J.J. Furedy, *Theories and Applications in the Detection of Deception* (Springer-Verlag, New York, 1990). The more pro-PDD side of the debate is typified by J.C. Kircher, S.W. Horowitz, and D.C. Raskin, Meta-analysis of mock crime studies of the control question polygraph technique, *Law and Human Behavior, 12*, 79–90 (1988); and D.C. Raskin, Polygraph techniques for the detection of deception, in *Psychological Methods in Criminal Investigation and Evidence*, D.C. Raskin, Ed. (Springer, New York, 1989).

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

5. G.H. Barland, C.R. Honts, and S.D. Barger, *Studies of the Accuracy of Security Screening Polygraph Examinations* (Research Division, Department of Defense Polygraph Institute, Fort McClellan, AL, 1989). Although this report and the results are unclassified, the government has limited its distribution and its first author has been forbidden by his parent agency from presenting the results at scientific meetings or from publishing the results in the scientific literature. Copies of the report are available from the present author.

6. These results are in sharp contrast to mock-crime studies of forensic PDD examinations, which often account for more than 70% of the criterion variance even on independent evaluation; see Kircher et al., note 3. The identity of the agency that produced better-than-chance results is classified.

7. Several follow-up studies have been conducted, and all have produced results consistent with the conclusion that national security polygraph tests produce large numbers of false negative errors. For example, see C.R. Honts, Counterintelligence scope polygraph (CSP) test found to be poor discriminator, *Forensic Reports, 5*, 215–218 (1992).

8. S.W. Horowitz, D.C. Raskin, C.R. Honts, and J.C. Kircher, *The directed lie: Standardizing control questions in the physiological detection of deception*, manuscript submitted for publication (1994). For both types of directed-lie questions, the subject agrees to the truthful answer, but is instructed to answer with a lie during the examination. The relevant-irrelevant test contains no items that are expected to elicit a response from innocent subjects and is now generally considered to be an invalid test; see Honts and Perry, note 1.

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

10. There are two principal reports: J.C. Kircher and D.C. Raskin, Human versus computerized evaluations of polygraph data in a laboratory setting, *Journal of Applied Psychology, 73*, 291–302 (1988); D.C. Raskin, J.C. Kircher, C.R. Honts, and S.W. Horowitz, *A study of the validity of polygraph examinations in criminal investigation* (Grant No. 85-IJ-CX-0040, National Institute of Justice), unpublished manuscript, University of Utah, Salt Lake City (1988).

11. For a detailed description of the bootstrapping analysis, see C.R. Honts and M.K. Devitt, *Bootstrap decision making for polygraph examinations: Final report of DOD/PERSEREC Grant No. N0014-92-1794* (Psychology Department, University of North Dakota, Grand Forks, 1992).