# Exploring Controversies in the Art and Science of Polygraph Testing

*Polygraph testing has always been controversial, with a wide gulf separating practitioners from scientists with relevant expertise. Myths about its alleged objectivity and infallibility stem from a failure to recognize that the polygraph simply records physiological responses and that the validity with which deception can be diagnosed depends on the nature of the examination and its social context.*

JOHN RUSCIO

Polygraph or "lie detector" tests have been conducted extensively in the United States since the early twentieth century, with comparatively little popularity elsewhere. The technique has been used to aid in criminal investigations, civil disputes, and the screening of job applicants and current employees. Polygraph testing has been controversial since its inception, with a substantial gulf separating practitioners from scientists with expertise in psychophysiology, measurement, decision making, and related disciplines. Aside from a trend toward computerized recording and scoring of polygraph charts, there has been little fundamental change in the technology or its implementation in the past fifty years, and a number of widespread myths persist despite

the lack of noteworthy advances in the field. Among the lay public, the press, and policy makers, confusion about the alleged objectivity and infallibility of polygraph testing often stems from a failure to recognize that the polygraph is simply a device to record physiological responses and that the validity with which deception can be diagnosed from such responses depends on the nature of the examination that is performed and its social context. My goal in this article is to dispel some of these myths by describing the different types and uses of polygraph tests, as well as empirical findings and additional considerations bearing on the validity and utility of those tests.

## Two Pervasive Myths of Polygraph Testing

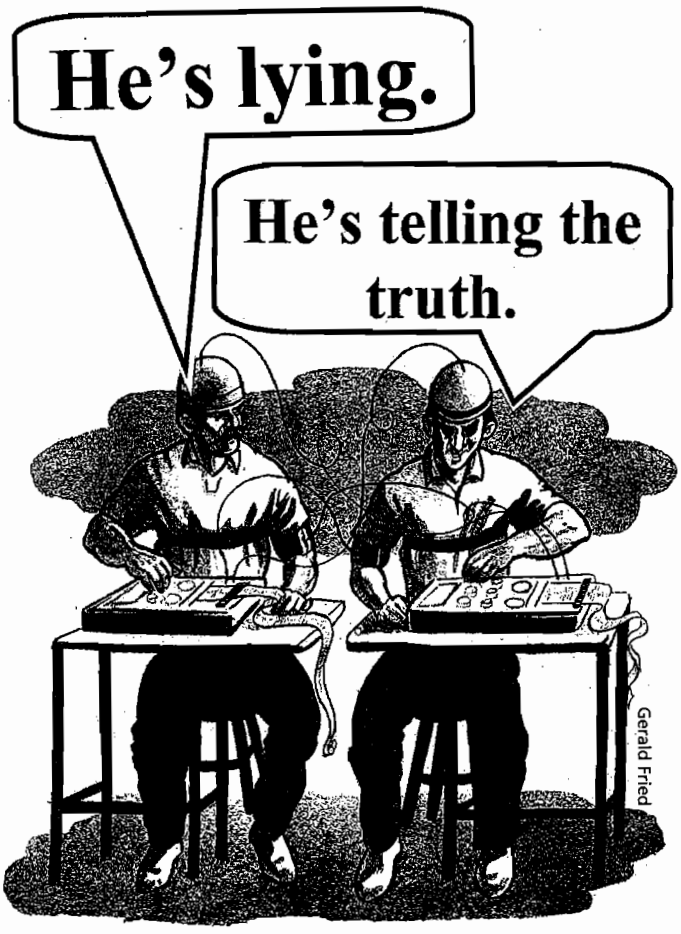
Many people perceive polygraph testing to be an *objective* way to distinguish honesty from deception, a scientific practice that is carried out in an impartial manner. The gadgetry of the standard polygraph device—which measures electrodermal, cardiovascular, and respiratory responses—may be responsible for this scientific mystique, yet with the exception of one type of application, polygraph testing has been developed and promoted primarily by nonscientists (Lykken 1998). A polygraph examination involves a lengthy pretest interview during which the examiner gains access to a wealth of background information and formulates specific questions that are then used during the test phase, when physiological responses are recorded. The type of questions that are constructed and the interpersonal manner with which they are posed can have a significant influence on the examinee's physiological responses. Moreover, the scoring and interpretation of the polygraph chart also require subjective judgments that can be influenced by an examiner's preconceptions, and there remains a critical need for well-validated computerized scoring and interpretive systems.

An exposé on private polygraph firms aired by *60 Minutes* in 1986 underscores the subjectivity of the process and its susceptibility to confirmation bias due to the contamination of an examiner's diagnosis with information obtained outside of the formal polygraph test. As summarized by Ben-Shakhar (1991, 236):

> Three different polygraph firms were independently called to test an alleged theft of a camera and lens from a photography magazine office employing four employees. In fact, nothing was stolen from the office, but the polygraph examiners were told that it could only have been done by one of the four employees. Each polygraph examiner was told that "it might have been ____," with a different employee being fingered in each case (a decidedly weak fingering). In each case, the polygraph examiner identified the "fingered" employee as deceptive, and cleared the other "suspects." Moreover, all polygraph examiners expressed complete confidence in their decisions. This demonstrates not only that polygraph examiners can go wrong, but that their judgment and decision-making processes are infected by a systematic and powerful source of bias, a bias caused by contamination.

In addition to potential biases in the administration and interpretation of a polygraph test, it is important to note that there are no unique physiological indices of deception (Lykken 1998; National Research Council [NRC] 2003). A number of psychological states influence the processes measured by the polygraph. For example, even innocent suspects may be alarmed or aroused by questions that imply wrongdoing, and the fact that examinees recognize that they may be fined, fired, imprisoned, or otherwise sanctioned based on their responses to key questions can often cause them to exhibit many of the behavioral and physiological reactions often believed to be uniquely indicative of deception. In its review of the polygraph literature, the NRC (2003) faulted proponents for failing to establish a plausible theoretical base linking polygraph responses to deception or to draw from relevant basic science research to measure those physiological processes that are likely to yield the greatest validity. For example, certain measures of brain activity show considerable promise as physiological responses that are less susceptible to countermeasures because of their rapidity and automaticity, yet there is little ongoing research to suggest a serious interest in improving polygraph technology (NRC 2003).

A second myth involves alleged infallibility of polygraph testing, the notion that the technique can be used to distinguish honesty from deception without error. Proponents derive much of their confidence from understandable but seriously flawed inferences. For example, Lykken (1998) cites a statement made by one polygrapher who testified before the Minnesota legislature in 1975 that he'd given more than 20,000 polygraph tests and had "never been shown to have made a mistake" (69). Because polygraph tests are ordinarily administered when other evidence is weak, one may never know with reasonable



certainty whether the examinee was in fact guilty of the alleged misconduct; that is, "ground truth" remains unknown. To the extent that polygraph results are taken to be definitive, further investigation is seldom undertaken to determine the validity of the inferences drawn from polygraph tests.

Another flawed source of intuitively compelling support stems from the fact that many examinees make confessions or offer damning admissions after an examiner suggests that their responses were deceptive. Such confessions are insufficient to establish the validity, much less the infallibility, of the polygraph test itself for at least two reasons. First, many confessions themselves are false statements, perhaps offered in an attempt to mitigate the consequences of having failed the polygraph test. In some instances, examinees' own faith in the infallibility of the polygraph has caused them to doubt the accuracy of their own memory, supposing that they must be guilty of the alleged offense even if they cannot recall it (Lykken 1998).

A second problem with using confessions as the criterion against which to evaluate the validity of polygraph testing is that this introduces a substantial bias. A study by Murray (1989) nicely illustrates this flawed logic. Of the 552 actual criminal suspects given polygraph tests, 239 (43%) were diagnosed as deceptive and were pressed for confessions; 104 (44%) of these individuals confessed, and 3 other individuals were later cleared by the confessions of other suspects. The 313 suspects diagnosed as nondeceptive were not pressed for confessions, but 18 (6%) of them were cleared by subsequent confessions of others. Noting that only 3 errors had been documented (the individuals diagnosed deceptive who were later cleared by others' confessions), Murray reported an accuracy of 99.4% for the polygraph. This conclusion is entirely unwarranted, as it presumes that the remaining 132 individuals diagnosed deceptive who did not confess were in fact guilty *and* that the remaining 295 individuals diagnosed honest who did not confess were in fact innocent, yet there was no independent corroboration of *any* individual's guilt or innocence. To see just how illogical Murray's conclusion is, consider the fact that equal (or perhaps even superior) results may have been obtained by tossing a coin to diagnose deception rather than actually scoring and interpreting the polygraph charts. Roughly half of all suspects would be diagnosed deceptive by the coin toss, and a substantial proportion would no doubt have confessed (keeping in mind that the base rate of guilt in this sample was quite high). Likewise, among the half of all suspects diagnosed honest by the coin toss, some would later be cleared by the confessions of others.

The proper way to calculate the empirical accuracy of polygraph testing is to establish ground truth with reasonable certainty and tally the results for tests administered to *all* examinees, not simply the subsample whose confessions support the initial diagnosis of deception (even if purely by chance). Likewise, it is essential that those who score and interpret the polygraph charts remain blind to outside information relevant to the case. Otherwise, there is no way to determine whether examiners are making diagnoses on the basis of case facts available without the polygraph test or on the basis of the polygraph chart itself.[1] To date, no field study has been conducted that meets even these minimal standards of scientific objectivity.

So just how accurate are polygraph tests? In light of the present evidence, it is extremely difficult to reach a defensible judgment and virtually impossible to summarize the results with a single numerical estimate. The National Research Council (2003) conducted a thorough review of the empirical literature on polygraph testing. Although they found the internal and external validity of the studies performed to date to be generally quite weak, especially for field studies, they concluded that polygraph testing can achieve results significantly better than chance, but far from perfect. Thus, despite decades of widespread use and ample opportunity to conduct informative laboratory and field studies, the quality of polygraph research remains low and claims of infallibility or near infallibility remain unsupported.

## Uses and Types of Polygraph Tests

Broadly speaking, polygraph tests are used for two primary purposes: in investigating specific incidents (e.g., crimes, security violations) and for screening (e.g., evaluating the honesty of prospective or current employees). Although there are subtle variations, two primary types of polygraph tests are used: the control question test (CQT; Raskin and Honts 2002) and the guilty knowledge test (GKT; Lykken 1959; Nakayama 2002; also known as the "concealed information test"). I will explore in turn the use of each type of test for each purpose.

### Specific Incidents

When an examinee is suspected of having engaged in a particular act of misconduct, polygraph testing can be performed using either type of test. For a CQT, the examiner constructs a series of questions that includes items relevant to the specific incident (e.g., "Did you steal $2,000 from your employer?") and items intended to serve as a point of comparison (e.g., "Since the age of eighteen, have you ever lied to get out of trouble?"). The latter questions are designed to be psychologically arousing to most examinees, with the presumption that anyone would need to lie (rather than overtly admit to past dishonesty, which would be perceived as incriminating in its own right) because of the ubiquity of at least minor acts of dishonesty in everyone's past. Such items are often referred to as "control" questions, although they do not provide reasonable controls in the scientific sense of the term.

The rationale underlying the interpretation of a CQT is that guilty individuals will be more aroused by the relevant than the control items, whereas innocent individuals will be more aroused by the control than the relevant items. Unfortunately, the presumed directionality of these responses reflects a naïve understanding of psychology. Because the relevant and control questions are discussed and agreed upon by the examiner and

*John Ruscio, Ph.D., is an associate professor of psychology at Elizabethtown College, One Alpha Drive, Elizabethtown, PA 17022. His research interests are in the areas of statistical methods for classification research and the psychology of judgment and decision making. His e-mail address is rusciojp@etown.edu.*

examinee during the pretest interview, they are easily distinguished during the test phase. This has two consequences, both of which tend to undermine the validity of the CQT. First, both innocent and guilty suspects recognize which questions are important for their futures, and are likely to feel considerably more threatened by the relevant questions. Not surprisingly, many studies of the CQT reveal an alarmingly high rate of false-positive diagnoses of deception: It is not uncommon for 40 percent to 50 percent of innocent suspects to fail a CQT polygraph examination (Lykken 1998; NRC 2003). Second, whether innocent or guilty, savvy examinees can take advantage of any number of countermeasures to systematically increase their physiological responses to the control questions, thereby passing the test. Lykken notes that laboratory studies have shown that college students can be trained to beat this type of polygraph test in under an hour. One can only suppose that individuals facing serious consequences on the basis of real-life polygraph tests would be especially motivated to learn and practice countermeasures prior to their examinations.

Whereas the CQT is based on several implausible assumptions (Lykken 1998) and nonetheless predominates in practice, the GKT has been developed by scientists and holds much greater promise for the investigation of specific incidents. Rather than attempting to determine the truthfulness of an examinee's responses to relevant items, this technique aims to assess whether an examinee possesses knowledge to which only a guilty individual would have access. To do so, a series of multiple-choice questions is constructed such that one choice is correct and the others are equally plausible to an innocent suspect. For example, one item relevant to a murder investigation could include a number of potential murder weapons, another could include descriptions of the victim's clothing at the time of the attack, and so forth. A total of six choices are usually provided for each question, with the correct alternative positioned randomly among the latter five choices (due to primacy effects, the first choice is not scored). The examinee is asked to repeat each of these choices out loud, which may serve to thwart potential countermeasures based on self-induced competing thoughts, and a score is calculated based on the frequency with which the maximal physiological response coincides with the correct alternative. An examinee who consistently responds most strongly to the correct choices is judged to possess knowledge of the incriminating information, whereas an examinee who responds to the correct choices at chance levels is judged not to possess this knowledge.

The larger the number of good items, the more accurately one can draw conclusions from a GKT. A *good item* is one for which the correct choice is salient to a guilty individual, but all choices are equally plausible to innocent individuals. One can pilot-test candidate GKT items on known-innocent examinees to evaluate the latter criterion, but the former may be more difficult to establish. The guilty individual must have noticed the relevant stimulus at the time of the incident and remember it during the polygraph test. To the extent that a large number of good items can be constructed, a GKT can differentiate those who do and do not possess the incriminating information with impressive accuracy (Beh-Shakhar and Elaad 2003). In stark contrast to the CQT, it is noteworthy that the GKT provides excellent protection for innocent suspects because the rate of false-positive identifications can be minimized by using a large number of items and setting a stringent threshold for diagnosing the possession of guilty knowledge. Field studies of the GKT are badly needed, but for reasons I'll discuss shortly law enforcement agencies have been uncooperative with scientists who wish to perform such studies.

The rationale underlying the GKT is based on the well-established nature of an "orienting response." This is similar to the so-called "cocktail party effect" whereby one's attention is automatically drawn to familiar information, such as hearing your own name spoken at a crowded cocktail party. It is especially difficult to suppress such a response to familiarity, and it would be difficult to systematically increase one's physiological response to all of the incorrect alternatives. To do so would require a large number of self-induced increases and decreases in arousal. This would be difficult to effect repeatedly and would lead to two additional problems. First, employing countermeasures a great many times may tip off a skilled examiner, thereby signaling the intention to deceive. Second, it may be possible to detect statistical differences in physiological responding if the *smallest* response consistently coincides with the correct alternative or if far *fewer* maximal responses coincide with the correct alternatives than would be expected by chance. In other words, the successful use of countermeasures requires that one establish a genuinely chance-level pattern of responding, which would be exceedingly difficult.

Although detractors have argued that the GKT may be applicable to only a small proportion of specific incident investigations, this remains an open question. For example, if police investigators began collecting information specifically for the construction of a GKT as soon as they arrived on the scene of a crime, and if they successfully withhold information from the news media (details incidental to the story, yet salient to the perpetrator, could feasibly be kept secret), it may be possible to use the GKT quite often. In fact, the technique is being used regularly in Japan (Nakayama 2002). It appears that the primary obstacle to implementing the GKT in the United States is that police officers and the polygraphers they employ are too convinced of the validity of the CQT. Another potential problem: Whereas the CQT was developed and has been promoted by the law enforcement agencies, the GKT was developed and is widely endorsed by psychological scientists. Thus, an unfortunate turf battle may contribute to the reluctance to perform field studies of the GKT in the U.S.—scientists need the cooperation of the police to conduct such research. Finally, a large-scale switch from the CQT to the GKT may represent an admission that law enforcement agencies have long been using a flawed technique, and this may also call into question convictions that were based in part on polygraph evidence.

### Screening

Polygraph testing is also used to screen prospective or current employees. Under certain circumstances, it may be possible to use the GKT to screen employees suspected of committing a particular misdeed. In such "focused screening" applications, polygraph testing could proceed as it does in the context of

specific incident investigations. In the majority of screening applications, however, no particular transgression is suspected.

Pre-employment screening and either regular (e.g., every five years) or aperiodic (i.e., random time intervals) employee screening is performed to assess the honesty of employees without reference to any alleged or suspected specific incident in order to predict the likelihood that an examinee will perform acceptably on the job. Given the substantial problems of employee theft, sabotage, security violations, or other costly or dangerous misdeeds, past misconduct or dishonesty may suggest a risky hire or the need to terminate an employee. Because no specific incident is involved in these types of screening, the GKT cannot be used, and tests analogous to the CQT must be constructed. Physiological responses to questions pertaining to a number of different areas—such as general honesty or past misconduct such as theft or drug use—serve as points of comparison for one another. The use of more general questions likely decreases the magnitude of any differences in responding, and such differences are interpretationally ambiguous in any event because questions about some topics may be inherently more threatening or arousing to particular examinees for any number of reasons. Both of these problems are likely to substantially weaken the validity of such screening tests (NRC 2003).

Despite decades of use by private employers and government agencies, no field studies have evaluated the accuracy of screening tests in a manner that meets minimal scientific research standards (NRC 2003), and their use is therefore based on unsupported presumptions of utility. There is also no evidence to suggest that prior misdeeds are in fact useful predictors of future misconduct. For example, admissions of illegal drug use are commonly elicited by polygraph screening, but there is no evidence bearing on the extent to which they predict the future thefts, sabotage, or security violations that the polygraph screening is intended to prevent.

## Further Difficulties of Polygraph Testing

Particularly in screening contexts, there may be an extremely low *base rate* of deception, which mathematically guarantees a large number of false-positive test results. For example, in the event that classified intelligence information is shared with a foreign power, there may be many individuals who had access to the information, only one of whom engaged in espionage. Screening all available suspects will inevitably yield many false accusations, and may in fact fail to identify the guilty individual. Altering the threshold for diagnosing deception presents policy makers with a choice between two unpleasant alternatives. On the one hand, setting a low threshold will increase the chances of correctly diagnosing deception in the examination of the guilty individual, yet it will also increase the number of false-positive results. All those who fail their tests will be indistinguishable by the polygraph alone, necessitating a large number of thorough investigations. As the NRC report notes, in the context of a low base rate and an assessment tool with as poor specificity as polygraph tests, it is easily possible that one might falsely implicate *hundreds* of innocent individuals in order to potentially identify the lone culprit. An agency may not have the resources to perform the required follow-up investigations, and there are a number of problems with revoking security clearances for all those who fail their tests. On the other hand, setting a stringent threshold will reduce the number of false-positive polygraph results, but at the cost of decreasing the odds of identifying the actual culprit and thereby defeating the purpose of the screening. Clearly, neither option is appealing, and the only way to improve matters is to develop an investigative tool more valid than polygraph testing or to restrict its use to situations in which the base rate of deception is not prohibitively low.

Another challenge for polygraph testing involves the high-stakes nature of many situations in which one wishes to diagnose deception. For example, a spy may be highly motivated to learn and practice effective countermeasures in anticipation of subsequent polygraph screening. Convicted spy Aldrich Ames passed several polygraph screening tests during his career at the CIA, which may have allayed suspicions of his misconduct far longer than would otherwise have been the case. Thus, when polygraph screening is used to clear individuals of suspicion, normal security precautions may be relaxed due to the false sense of security provided by the test results. The NRC highlighted this danger and advised caution in the use of polygraph testing for security screening. In contrast, innocent suspects may falsely believe in the validity of the polygraph and be less likely to engage in countermeasures and therefore more susceptible to failing polygraph tests. Or, the scrupulously honest individuals that one would wish to hold top-level security clearances may fear false-positive polygraph results and the damaging effects on their careers and choose to pursue other lines of work. Agencies that use polygraph testing for security screening purposes have suffered immeasurably from the flight of qualified employees who fear the repercussions of false accusations as well as an accompanying loss of morale among the employees who remain (NRC 2003).

## Conclusions

As any proponent will quickly attest, polygraph testing has proven useful in eliciting confessions and admissions of wrongdoing. This provides considerable support for the utility of such testing as an investigative tool even if the tests themselves are of imperfect validity. However, such benefits are offset by a number of other weaknesses that suggest caution in drawing conclusions from any polygraph tests that fail to produce such confessions. Moreover, it is essential to verify the accuracy of a confession elicited by a polygraph test.

In 1993, the U.S. Supreme Court established criteria for the admissibility of scientific evidence, including general acceptance of the relevant theory and technique within the appropriate scientific discipline, a known or probable error rate that is satisfactory, standards that maximize the validity of a technique in practice, and publication in peer-reviewed scientific journals (*Daubert* v. *Merrell Dow Pharmaceuticals*, 1993). Whereas a strong case has been made for the inadmissibility of CQT results as evidence in court under the *Daubert* guidelines (Saxe and Ben-Shakhar 1999), pending successful field validation an equally strong case may be made for the admissibility

of GKT results (Ben-Shakhar, Bar-Hillel, and Kremnitzer 2002). Interestingly, most states prohibit admission of polygraph evidence in civil and criminal cases, but many individuals foolishly stipulate to the admission of test results. For example, in a criminal case that a district attorney wishes to close due to a lack of sufficient evidence or promising leads, a deal may be offered that many defendants find attractive: The charges will be dismissed if the suspect passes a polygraph test, but the test results will be admissible in court in the event that the suspect fails. Believing in the validity of the test and hoping to bring the ordeal of pending prosecution to an end, many suspects accept such an offer and are then surprised when they fail the test and thereby provide the state with evidence that many jurors find persuasive. Likewise, disputants in civil suits (e.g., both parents vying for custody of their children in a divorce) will sometimes agree to each take a polygraph test, with both sets of results stipulated in advance to be admissible. This, too, often turns out to be a foolish bargain. One can reasonably wonder whether other evidence of dubious validity should be allowed into the court record merely because both parties stipulate to it in advance (e.g., could the toss of a coin be used instead?).

In 1988, the federal Employee Polygraph Protection Act prohibited most private employers from using polygraph testing as a screening tool. Private firms, often the same ones that formerly administered large numbers of polygraph tests, have responded to fill this niche by offering paper-and-pencil honesty tests or voice stress analyzers that may be of even poorer validity (Lykken 1998; NRC 2003). However, it is curious that law enforcement, military, and government agencies were exempted from the prohibition on polygraph testing: A technique judged unsuitable for screening applicants to the local convenience store is still used to screen individuals responsible for maintaining our national security. It is time to carefully weigh the costs and benefits of polygraph screening in each particular context in which its use is being contemplated. The advantage of eliciting confessions and damning admissions must be considered against disadvantages such as invasions of privacy, damage to employee morale, the cost of investigating large numbers of false-positive results and the accompanying harm that is done to qualified individuals' careers, the possibility that the most dangerous individuals may be likely to learn and effectively employ countermeasures, and the false sense of security that can stem from the use of a highly fallible screening tool.

In its report prompted by the Department of Energy, the NRC offered recommendations that emphasize a number of steps to improve the practice of polygraph testing and safeguard our national security. The report encouraged research on the basic science relevant to the psychophysiological detection of deception such as the polygraph. Alternative processes and techniques that can supplement or replace the polygraph should receive serious study. This research should not be conducted by agencies with a real or perceived conflict of interest. For example, the Department of Defense Polygraph Institute is currently responsible both for the development and evaluation of polygraph techniques as well as training practitioners and promoting the use of the polygraph. Open scientific communication of research is also important, with results to be classified only when absolutely necessary to maintain national security. Above all, polygraph results should be interpreted with far greater caution than is currently evident, and policy makers should pay increased attention to the inherent tradeoffs of false-positive and false-negative identifications that stem from shifts in the threshold used to diagnose deception. The research program outlined above, along with a proper recognition of the strengths and limitations of the psychophysiological detection of deception, is essential to undertake a realistic assessment of the utility of polygraph testing.

## References


Ben-Shakhar, Gershon. 1991. Clinical judgment and decision-making in CQT-polygraphy: A comparison with other pseudoscientific applications in psychology. *Integrative Physiological and Behavioral Science* 26: 232–240.

Ben-Shakhar, Gershon, Maya Bar-Hillel, and Mordechai Kremnitzer. 2002. Trial by polygraph: Reconsidering the use of the guilty knowledge technique in court. *Law and Human Behavior* 26: 527–541.

Ben-Shakhar, Gershon, and Eitan Elaad. 2003. The validity of psychophysiological detection of deception with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology* 88: 131–151.

*Daubert v. Merrell Dow Pharmaceuticals, Inc.* 1993. 509 U.S. 579.

Lykken, David T. 1959. The GSR in the detection of guilt. *Journal of Applied Psychology* 43: 385–388.

———. 1998. *A Tremor in the Blood: Uses and Abuses of the Lie Detector.* New York: Plenum Trade.

Murray, Kenneth E. 1989. Movement recording chairs: A necessity? *Polygraph* 18: 15–23.

Nakayama, Makoto. 2002. Practical use of the concealed information test for criminal investigation in Japan. In Murray Kleiner (Ed.), *Handbook of Polygraph Testing.* San Diego: Academic Press.

National Research Council. 2003. *The Polygraph and Lie Detection.* Washington, D.C.: National Academies Press.

Raskin, David C., and Charles R. Honts. 2002. The Comparison Question Test. In Murray Kleiner (Ed.), *Handbook of Polygraph Testing.* San Diego: Academic Press.

Saxe, Leonard, and Gershon Ben-Shakhar. 1999. Admissibility of polygraph tests: The application of scientific standards post-Daubert. *Psychology, Public Policy, and Law* 5: 203–223.


## Note

1. Even this precaution does not guard against the possibility that examiners may construct or pose questions in a manner that biases physiological responses toward apparent deception or honesty in accordance with their preconceptions, but it does represent a minimally acceptable scientific procedure for evaluating the validity of conclusions drawn from polygraph tests. □