CHAPTER 19

Employing Polygraph Assessment

WILLIAM G. IACONO AND CHRISTOPHER J. PATRICK

ALTHOUGH the exact number is not known, it is a safe bet that tens of thousands of polygraph tests are administered in the United States every year. Most of these tests are administered by federal agencies as part of the government's national security screening program, and some are given by law enforcement to screen the integrity of potential new recruits. A substantial fraction are forensic polygraphs administered by law enforcement as an investigative tool to assist the resolution of criminal cases. Some come from criminal defendants who hire examiners in private practice with the hope of obtaining exculpatory outcomes. Others arise from civil cases involving parental custody/fitness, sex offender commitment, and employee rights. In any case, it is unlikely that a forensic psychologist has administered a polygraph. Instead, polygraphs are administered by polygraphers who work in a profession that is largely disconnected from psychology and informed little by psychological science. Our aim in this chapter is to bridge this gap between applied polygraphy and forensic psychology by providing the information needed to critically evaluate polygraph practice. In addition to examining the current state of polygraph testing, we also review future possible applications of deception detection techniques.

CURRENT APPLICATIONS

Conventional polygraph tests typically are used when the question at hand cannot be easily resolved by the available evidence. When the investigation reaches an evidentiary dead end, police may rely on a polygraph test of a known suspect as the means of last resort to resolve the case. Sometimes those who fail these tests, pressured to own up to their misdeeds, confess, thereby providing the police with incriminating evidence they otherwise would not have. In the absence of a confession, a failed test may lead the police to cease the investigation, believing the suspect at hand is guilty even if the evidence is insufficient for successful prosecution. By contrast, a passed test provides incentive to continue the investigation and look for new suspects.

Polygraph tests are relied on by psychologists in a number of ways:

- In sex offender treatment programs to ensure that offenders are fully disclosing their offenses and fantasies
- By insurance agencies to verify the claims of those insured
- In family court to help resolve charges of misbehavior parents level at each other in their effort to obtain custody of their children
- By the police to verify victims' charges
- By controversial people in the public eye who wish to sway public opinion in their favor by advertising the fact that they passed a "lie detector"
- By the government to protect national security by requiring those with access to classified information to pass tests confirming that they are not spies
- Even by those running fishing contests to verify that winners actually followed contest rules rather than purchasing their lunker from the local supermarket

The Employee Polygraph Protection Act (EPPA; 1988) eliminated much of the most widespread application of polygraph testing, the periodic screening of employees to verify their good behavior and the pre-employment screening of potential hires to see if they possess the qualities desired by the employer. Ironically, the government exempted itself from coverage by this law and has been expanding polygraph testing programs in light of concerns about terrorism and national security. For instance, since the passage of the EPPA, Public Law 106-65, passed as part of the National Defense Authorization Act (2000), requires scientists at nuclear weapons laboratories to submit to polygraph tests to maintain their security clearance. Besides many state and local law enforcement agencies and polygraphers in private practice, over two dozen federal agencies routinely use polygraph tests, including those that are part of the Departments of Defense, Energy, Homeland Security, Justice, and Treasury.

THE POLYGRAPH AND THE POLYGRAPH EXAMINER

Traditional polygraphs are briefcase-size instruments that use moving chart paper to record the autonomic responses elicited by the subject's answers to test questions. Although these devices are still in use, portable computers that digitally record autonomic activity, displaying and storing it in a manner that mimics the appearance of paper chart recordings, are now in common use. Expandable pneumatic belts positioned around the upper thorax and abdomen provide two separate recordings of the chest movements associated with inspiration and expiration. Changes in palmar sweating (skin conductance, aka the galvanic skin response [GSR]) are detected by electrodes attached to the fingertips. For the "cardio" channel, a partly inflated blood pressure cuff attached to the arm reflects relative changes in blood pressure and provides an index of pulse. Occasionally a fifth channel monitoring blood flow to the fingertip is included. Although this instrumentation is relatively simple, it produces valid records of physiological reactivity that are comparable to those obtained by sophisticated laboratory equipment (Patrick & Iacono, 1991a).

Training in polygraphy is provided by free-standing polygraph schools, most of which are accredited by the American Polygraph Association. The most prestigious of these is at the National Center for Credibility Assessment (NCCA; formerly the Department of Defense Polygraph Institute) located at Fort Jackson, South Carolina. This school offers a one-semester, intensive, hands-on course in polygraphy that covers ethics, law, the physiology and psychology of deception detection, and the various techniques and interview practices employed by examiners. Graduates of the program typically are apprenticed to practicing examiners before becoming fully certified to administer tests on their own. NCCA offers training for many state and city police departments and most federal government agencies, including the military police, the Federal Bureau of Investigation, the Internal Revenue Service, and all of the government security agencies. NCCA also has an in-house research program staffed by doctoral-level psychologists, some of whom share in the teaching of students with polygraph examiners and law enforcement agents.

NCCA, which requires a college degree and two years of law enforcement experience for program admission, represents the best training the profession of polygraphy has to offer. Most accredited schools do not offer as rigorous a program; not all practicing polygraph examiners are graduates of approved schools; and, because polygraphy is not regulated in most states, polygraphers are not necessarily licensed to practice their trade.

POLYGRAPH TECHNIQUES

The polygraph instrument is not capable of detecting lies, and no pattern of physiological response is unique to lying. Consequently, all polygraph techniques involve asking different types of questions, with differential responding to those pertinent to the issue at hand determining outcome. The techniques, all of which have multiple variants, fall into two categories involving either specific incident or personnel screening applications.

Specific Incident Investigations

There are three types of specific incident polygraph tests. These procedures are applied when polygraph examiners are aware that an event has occurred but are uncertain what role the examinee played in the incident.

Control (Comparison) Question Technique. The so-called control or comparison question technique (CQT) remains the procedure of choice for specific incident investigations like those concerned with known criminal acts. The CQT typically consists of about 10 questions. The two types of question that are important to the determination of guilt or innocence are referred to as relevant and control questions. The relevant questions deal directly with the incident under investigation (e.g., Did you shoot Bill Birditsman on the night of March 18?). Control items are paired with relevant questions and cover past behaviors that one might associate with "the kind of person" who is capable of killing (e.g., Before the age of 24, did you ever deliberately hurt someone you were close to?). It is assumed that guilty suspects will be more concerned with the relevant questions than with the control questions. The reverse pattern is expected with innocent people.

The typical CQT has three parts: (1) a pretest interview (lasting between 30 minutes and 2 hours) during which the question list is formulated, (2) the presentation of the question list (usually repeated three times with the question order varied for each of the three "charts") while physiological responses are recorded, and (3) a posttest interrogation.

The pretest interrogation is designed to determine if the examinee is suitable for testing—for example, if he or she slept the night before and is in reasonably good health. It also provides an opportunity for the examinee to provide an account of the facts in dispute, information that is used in combination with the background material provided the examiner about the case to develop the test questions.

The pretest phase of the CQT is critical to the successful administration of the test. It is during this interview that the polygrapher attempts to create circumstances that lead the innocent person to be more disturbed by the possibly trivial issues raised by the control items than by the relevant questions that have to do with the matter under investigation. A common criticism of the CQT is that it is biased against truthful persons, because the relevant questions may be just as arousing to innocent suspects, who may view their freedom or livelihood as dependent on their physiological response to these items, as they are to the guilty (Lykken, 1974). To reduce the likelihood of this occurrence, polygraphers use the pretest interview to focus the subject's "psychological set" on the control questions if the examinee is innocent or on the relevant questions if she or he is guilty. Two tactics are used to accomplish this objective.

The first is to convince the subject that lies will be detected. One way to achieve this goal is to demonstrate that the polygraph can detect a known lie. In a typical scenario, the examiner connects the subject to the polygraph and says, "I'm going to ask you to pick a number from 1 to 10, write it down, and then show it to me. Both of us will know which number you've picked. After that, I will say a number and ask you if it is yours. I want you to answer 'no' to each number I say, including the one you picked." The examiner then records the subject's responses to each number and tells him or her afterward that the largest reaction occurred when the person lied; if this was indeed the case, the examiner may point it out to the subject on the chart. If it was not the case, the examiner may imply that it was anyway ("I can see from the results that I will be able to tell when you are lying or telling the truth") or alter the subject's response to the target number to create the impression that it elicited a clearly detectable reaction. Some examiners achieve the desired result by having the subject pick a card from a stacked deck and then rely on the physiological record to "determine" which one he or she picked. Most polygraphers routinely use some variant of this type of demonstration procedure, often called a stim or acquaintance test.

A second tactic for establishing the correct psychological set is to continually emphasize the importance of always being truthful. No distinction is made between the relevant and the control questions regarding the burden of truthfulness. Consequently, innocent individuals are led to believe that lying to control questions will lead to a failed test outcome. How it is that they should reach this conclusion is explained for a case of theft by one of polygraphy's leading proponents, David Raskin (1989), as follows:

> Since this is a matter of a theft, I need to ask you some general questions about yourself in order to assess your basic honesty and trustworthiness. I need to make sure that you have never done anything of a similar nature in the past and that you are not the type of person who would do something like stealing that ring and then would lie about it. . . . So if I ask you, "Before the age of 23, did you ever lie to get out of trouble . . . ?" you could answer that no, couldn't you? Most subjects initially answer no to the control questions. If the subject answers yes, the examiner asks for an explanation . . . and] leads the subject to believe that admissions will cause the examiner to form the opinion that the subject is dishonest and therefore guilty. This discourages admissions and maximizes the likelihood that the negative answer is untruthful. However, the manner of introducing and explaining the control questions also causes the subject to believe that deceptive answers to them will result in strong physiological reactions during the test and will lead the examiner to conclude that the subject was deceptive with respect to the relevant issues concerning the theft. In fact, the converse is true. Stronger reactions to the control questions will be interpreted as indicating that the subject's denials to the relevant questions are truthful. (pp. 254–255)

Q1: Please provide opening bracket for "...explanation. . . and]"

Charts are scored using one or a combination of three approaches. With global scoring, all the information available to the examiner is used to make the determination of truthfulness. Hence, in addition to inspection of the physiological data, the plausibility of the subject's account of the facts during the pretest interview, his or her demeanor during the examination, and information from the investigative file may all figure into the evaluation.

With now widely employed numerical scoring, the examiner derives a score from the physiological recordings. The magnitude of the response to pairs of control and relevant questions is estimated for each separate physiological channel. In the most commonly employed of several popular methods, a score from +1 to +3 is assigned if the response to the control item is larger, with the magnitude of the score determined by how large a difference is observed. Likewise, a score from −1 to −3 is assigned if the relevant item of the question pair elicited the stronger response. A total score is obtained by summing these values over all channels and charts, with a negative score less than −5 prompting a deceptive verdict, a positive score exceeding +5 a truthful verdict, and scores between −5 and +5 considered inconclusive and therefore warranting further testing. In our experience with government examiners, about 10% of CQTs end with inconclusive outcomes.

Both global and numerical chart evaluation have high interscorer reliability. Studies in which examiners blind to case facts evaluate the original examiners' charts typically report reliabilities around .90 (e.g., Honts, 1996; Horvath, 1977; Patrick & Iacono, 1991a, 1991b. The retest reliability of polygraph testing has not been evaluated. The absence of such data is unfortunate, because often questions about the possible increment in validity gained by retesting a defendant arise in legal hearings regarding the possible admissibility of polygraph results. In addition, the CQT, a collection of different techniques, is not a standardized test, so in the absence of retest data, it is not known to what degree examiners, all of whom have their own way of administering the CQT, are likely to obtain the same result when testing the same individual.

The third approach to chart scoring derives from computerized recording systems. Typically the computer provides a verdict in the form of a probability statement as to the likelihood the person was truthful when responding to the questions. Because these systems are marketed commercially, the algorithms and data used to justify the probability statements are proprietary. Although computer scoring is reliable, little is known about the validity of the outputted probability statements, and few polygraphers rely exclusively on computer scoring of charts, especially in forensic evaluations.

Once the charts are scored, the posttest phase of the CQT is launched. Those individuals who are believed to have been untruthful are interrogated during this phase. The point of the interrogation is to leverage the polygraph test outcome to obtain incriminating admissions or an outright confession. During this phase, skillful interrogators may resolve a case that otherwise would never have been resolved. It is this hoped-for outcome, which speaks to the utility and not the validity of the CQT, that keeps the CQT in widespread use despite its general inadmissibility as evidence in legal proceedings.

Directed Lie Technique. The directed lie technique (DLT) is considered a subtype of the CQT. The chief difference lies in the nature of the control questions. For a DLT, the "probable lie" control questions of the CQT are replaced with "directed lie" questions. Directed lies are statements that the subject admits involve a lie before the test begins. In fact, the polygrapher specifically instructs the subject to answer the question deceptively and to think of a particular time when he or she has done whatever the directed lie question covers. Examples of directed lies are "Have you ever done something that hurt or upset someone?" or "Have you ever made even one mistake?" As with the CQT, guilty subjects are expected to respond more strongly to the relevant questions, and innocent subjects should react more strongly to the directed lies.

Guilty Knowledge or Concealed Information Test. An alternative to the CQT for specific incident investigations is the guilty knowledge test (GKT; Lykken, 1959, 1960), sometimes referred to as a concealed information or knowledge test. Rather than asking directly whether the examinee was responsible for the crime under investigation, the GKT probes for knowledge indicative of guilt—details regarding a crime or incident that only the person who did it would know about. The GKT consists of a series of questions about the crime posed in multiple-choice format. Each question asks about one specific detail of the crime and is followed by a series of alternative answers, including the correct answer as well as other plausible but incorrect options. The following is an example of a GKT question concerning one detail of a homicide: "If you were the one who beat Donna Fisbee to death, then you will know what was used to kill her. Was she beaten with: (a) a brick? (b) a crowbar? (c) a pipe? (d) a baseball bat? (e) a hammer?" When presented with a question of this type, the true culprit would be expected to emit a larger physiological reaction to the correct alternative than an innocent person who knows nothing about the incident and would respond at random.

The simple premise underlying the GKT is that a person will exhibit larger orienting reactions to key information only if he or she recognizes it as distinctive or important. The GKT tests for knowledge of information rather than for deceptiveness, and the irrelevant alternatives are true controls rather than pseudocontrols. In the CQT, deceptiveness is inferred from a pattern of enhanced reactions to relevant questions, but the possibility that "innocent concern" rather than deception is responsible for this outcome can never be ruled out. A pattern of consistent reactions to critical items on a GKT can (within a small, estimable probability) mean only that the examinee possesses guilty knowledge. On a GKT question with five alternative answers, the odds that an innocent person with no knowledge of the crime would react most intensely to the key (relevant) alternative are 1 in 5. On a GKT that included 10 such questions, the odds are vanishingly small (<1 in 10,000,000) that an innocent person would react differentially to the key alternative on each and every test question.

The first study of the GKT (Lykken, 1959) and most others conducted since have utilized peripheral response measures, most commonly skin resistance or skin conductance, as indices of stimulus orienting. More recently, brain potentials recorded from the electroencephalogram have been utilized to detect deception within a GKT format. Measuring how reaction times differ to GKT key and irrelevant multiple choice alternatives has provided another method for identifying those with guilty knowledge (Seymour & Fraynt, 2009). The "attentional blink" paradigm has also been adapted to the GKT (Ganis & Patnaik, 2009). This paradigm makes use of the fact that when two stimuli are presented in close temporal proximity, attention to the first stimulus in the pair (which may or may not convey guilty knowledge) makes identifying the second stimulus difficult (causing a "blink" in attention).

PERSONNEL SCREENING

Modern screening tests differ from specific incident tests in that it is not known whether any particular transgression has taken place. Consequently, the relevant questions typically cover extended periods of time and many topics, leaving ambiguous what form an adequate "control" question should have. Whereas there are many different types of screening tests, these procedures are historically linked to the relevant/irrelevant technique (RIT), a polygraphic interrogation method that preceded the development of the CQT and was used originally in criminal investigations.

Relevant/Irrelevant Technique. In the original RIT, relevant questions (like those used on the CQT) were each preceded and followed by an irrelevant question (e.g., "Is your name Ralph?" or "Is today Tuesday?"). Consistently greater reactions to the relevant items of the test were interpreted as evidence of deceptiveness. However, because of the obvious confound posed by the differential potency of the two categories of questions, the traditional RIT has been roundly criticized and thus is used only occasionally today. For purposes of employment screening, polygraph examiners now commonly use a variant of the RIT procedure that might more appropriately be called the relevant/relevant technique, because interpretation of test outcome depends on the pattern of responses across all of the relevant questions.

In contrast to specific incident tests, screening examinations contain relevant questions of the form "Have you ever . . . ?" or "During the period in question, did you . . . ?" These questions, which may tap themes related to drug use, trustworthiness, and rule violations, are alternated with innocuous or irrelevant questions (also called norms). Law enforcement and security agencies use these types of tests both with prospective and current employees. Although government secrecy makes it difficult to determine how these two types of subjects fare on these tests, it is clear that prospective employees are much more likely to fail such tests (perhaps a third or more do, depending on the government agency) than those already screened, trained, and employed (where failure rates hovering around 1%–2% are seen).

In a screening test of this type, typically three or more question sequences are presented covering the same topics, but with the form of the questions and their order varied. The irrelevant items are included mainly to provide a rest period or return to baseline rather than a norm for comparison purposes. The RIT is a polygraph-assisted interview in which the development of questions is guided both by the polygrapher's impressions of the examinee's truthfulness as well as the comparative reactions to the various relevant items: "The cardinal rule in chart interpretation is, any change from normal requires an explanation" (Ferguson, 1966,

p. 161). If the subject shows persistently strong reactions to one or more content areas in relation to the rest, the examiner concludes that the subject lied or was particularly sensitive about these issues for some hidden reason. In this case, the examiner will probe the examinee for an explanation of what might have provoked these responses and will administer additional question sequences focusing on these specific issues. Examinees who are adept at explaining away their reactions are thus likely to avoid incrimination. Thurber (1981) reported that, among applicants for a police training academy, those who scored highest on a questionnaire measure of impression management were most likely to pass a polygraph screening test.

National security organizations use both periodic and aperiodic screening tests. Periodic screening tests are conducted at regular intervals to determine whether existing employees have been honest in their work and remain loyal to the agency. Aperiodic screenings are conducted less frequently and with minimal advance warning. Besides being more economical, this practice is thought to produce a more powerful deterrent to malfeasance. The knowledge that they may be asked to submit to a polygraph test at any time is believed to dissuade existing employees from engaging in misconduct. In effect, the polygraph establishes a climate of fear in which employees presumably are less inclined to be dishonest because they fear detection (National Research Council [NRC], 2003; Samuels, 1983).

Test for Espionage and Sabotage. In addition to RIT-derived tests, national security agencies have introduced a type of directed lie test as part of their counterintelligence program called the Test for Espionage and Sabotage (TES; or test for espionage, sabotage, and terrorism, TEST), a procedure that has been used extensively with scientists at nuclear weapons laboratories. With the TES, questions such as “Have you given classified information to any unauthorized person?” are paired with directed lies such as “Did you ever violate a traffic law?” Unlike other types of screening tests, the TES can be scored using the same procedures followed for the CQT.

DETERMINING VALIDITY

Hundreds of papers discuss the validity of polygraph testing. Much of this work is unpublished, and much that is published appears in poor-quality or trade journals. Because so many studies touch on the accuracy issue, and because much of the research conducted in this field is not carried out by scientists or published in scientific, peer-review journals, we preface our evaluation of the literature with a summary of the important methodological issues that a serious investigation of polygraph validity must address.

Evaluation of Polygraph Charts

Although currently semi-objective numerical scoring is the preferred technique for chart evaluation among professional polygraphers, the global approach to chart interpretation still is used occasionally. For CQTs conducted using either procedure, the field examiner is exposed to extrapolygraphic cues, such as the case facts, the behavior of the suspect during the examination, and sometimes inculpatory admissions from the examinee. For a validity study to provide a meaningful estimate of the accuracy of the psychophysiological test, the original examiner's charts must be reinterpreted by blind evaluators who have no knowledge of the suspect or case facts. Even though those trained in numerical scoring are specifically taught to ignore extrapolygraphic cues, Patrick and Iacono (1991b), in their field study of Royal Canadian Mounted Police (RCMP) polygraph practices, showed that even these elite examiners nevertheless attend to them. In 21% of the 279 examinations investigated, the original examiners contradicted the conclusions dictated by their own numerical scores by offering written verdicts that were not supported by the charts. We also found that original examiner opinions were likely to be more accurate than their numerical scores, indicating that examiners improved their accuracy when they relied on case facts and other extraneous information. Although one may be tempted to use such data to argue that blind chart scoring underestimates the accuracy of polygraph verdicts (e.g., see Honts, Raskin, & Kircher, 2002), the probative value of the CQT derives from the possibility that the psychophysiological measurements provide a scientifically valid method for detecting liars. No court of law would accept as evidence the opinion of a human "truth verifier," a skilled interviewer who can use the available evidence to reach a correct judgment. The fact that our RCMP data showed that original examiners were more accurate when they overrode the charts speaks to the invalidity of the psychophysiological test when used to determine truthfulness.

FIELD VERSUS LABORATORY INVESTIGATIONS

Field studies, like our study with the RCMP just discussed, involve real-life cases and circumstances. The subjects are actual criminal suspects. Laboratory studies require naive volunteers to simulate criminal behavior by enacting a mock crime. The latter approach provides unambiguous criteria for establishing ground truth but cannot be used to establish the real-life error rate, because the motivational and emotional concerns of the suspects are too dissimilar from those involved in real-life examinations. Unlike those faced with an actual criminal investigation, guilty subjects in the laboratory have little incentive to try and no time to research how to "beat" the test, guilty subjects are following instructions to lie rather than lying out of self-interest, and both guilty and innocent subjects have little to fear if they are classified as deceptive. Administering the CQT to laboratory subjects is especially likely to lead to overestimates of accuracy for the innocent. Innocent subjects can reasonably be expected to respond more strongly to the potentially embarrassing control questions concerning their personal integrity and honesty than to the relevant questions dealing with a simulated crime they carried out only to satisfy experimental requirements. However, laboratory research does permit efficient investigation of the influence of factors that may affect test outcome (e.g., effects of CMs or personality traits).

Laboratory studies of the GKT are also likely to overestimate its accuracy, more so for guilty than innocent individuals. Well-designed laboratory experiments construct a scenario in which guilty participants must attend to details of the “crime” that the examiner expects perpetrators to know and that can be used to construct the GKT. In real life, a criminal may not attend to the aspects of a crime that an investigator views as salient, and many details may be forgotten. For example, there is evidence that psychopathic individuals are less able to process to incidental details when focusing on a primary task (Kosson, 1988), and such individuals may thus be less detectable using the GKT (Verschuere, Crombez, De Clercq, & Koster, 2005; Waid, Orne, & Wilson, 1979). If a person does remember the details of a real-life crime, however, his or her recognition should evoke greater physiological reactions, thereby making it easier to detect the guilty.

Although the GKT is used in Israel and exclusively in Japan, there are two reasons why it is seldom used in real-life investigations in North America. First, there is a prevailing belief among field examiners that the CQT is virtually infallible (Patrick & Iacono, 1991b). Thus, there is no need to develop an alternative procedure, especially one that is more complicated to administer than the CQT. Second, to construct a valid GKT, there must be salient details of the crime known only to the perpetrator. Not all crimes meet this criterion, in part because often pertinent facts are generally known (e.g., through media reports). Rape provides a crime well suited for GKT development when the victim can provide pertinent crime details for test construction. Alleged sexual assaults in which the question of force versus consent is the only issue to be resolved would not be amenable to a GKT. However, DNA and fingerprint evidence are not available or necessarily relevant for many crimes, but this has not diminished their evidentiary value for those crimes where such evidence exists.

The problems with laboratory studies dictate that real-life applications must be used to evaluate polygraph tests. Although the CQT has been subjected to field research, there are no field studies of personnel screening tests and only two of the GKT, facts that limit efforts to evaluate these techniques.

Problems Establishing Ground Truth

The advantage of field investigations—that they are based on actual crimes—is also a significant drawback, because prima facie evidence of innocence or guilt is often lacking. Proponents of polygraphy have argued that confessions provide the best method for operationalizing ground truth. Confessions identify the culpable and clear the innocent. Although occasionally confessions are false, and those who confess may differ in important ways from those who do not, the major problem with this strategy concerns the likelihood that the confession is not independent of the original polygraph examiner’s assessment. For reasons that are unrelated to test accuracy, confessions are obtained during posttest interrogations and are associated almost exclusively with charts that indicate a deceptive outcome. When this occurs, the verified cases selected for a validity study will be biased in favor of demonstrating high accuracy for the technique.

To clarify this point, consider the following example. Ten women are suspects in a criminal investigation. A polygrapher tests them one by one until a deceptive outcome is obtained, say on the sixth suspect tested. (Under these circumstances, the remaining four women typically would not be tested, unless the crime was believed to involve more than one perpetrator.) According to usual practice, the examiner then attempts to extract a confession from the sixth suspect. If the examinee fails to confess, her guilt or innocence cannot be confirmed. It is possible that the polygrapher committed two errors in testing these six cases: The person with the deceptive chart may have been innocent, and one of those tested before her could have been guilty. In the absence of confession-backed verification, however, the polygraph records from these six cases will never be included as part of a sample in a validity study. If the sixth suspect does confess, however, these six charts, all of which confirm the original examiner's assessment, will be included. The resulting sample of cases would consist entirely of charts the original examiner judged correctly and would never include cases in which an error was made. As Iacono (1991) has shown, if polygraph testing actually had no better than chance accuracy, by basing validity studies on confession-verified charts selected in this manner, a researcher could misleadingly conclude that the technique was virtually infallible. Given how cases are selected in confession studies of validity, it should not be surprising that field validity studies typically report that the original examiner was 100% correct (or nearly so; see Honts et al., 2002) for the cases chosen for study. The case selection method assures this result.

Polygraph proponents have asserted that, because it is the original examiners who testify in court, it is the "accuracy" of the original examiners in these field confession studies that constitutes the "the true figure of merit" to determine how accurate polygraph tests would be in legal proceedings (Honts et al., 2002). Despite the fact that the hit rate of the original examiner in these studies is entirely misleading, given how cases are selected for study inclusion, this argument also ignores the contribution of extrapolygraph information to the original examiner's opinion and the resulting necessity of blind chart scoring to determine how useful the psychophysiological data are for deciding guilt.

WHAT CAN BE CONCLUDED ABOUT POLYGRAPH VALIDITY?

Different conclusions apply to the validity of each of the different types of polygraph procedures. Serious questions have been raised about the accuracy of each of the procedures that polygraph examiners commonly use. Ironically, the one procedure they seldom use, the GKT, has high potential validity.

Control Question Technique

The literature relevant to the validity of CQT polygraph testing has been reviewed repeatedly, including in the three prior editions of this text (Iacono & Patrick, 1987, 1999, 2006) as well as in other more recent publications (Iacono, 2007, 2008b, 2010; Iacono & Lykken, 2009; Meijer, Verschuere, Merckelbach, & Crombez, 2008; Vrij, 2008). Despite the importance of determining CQT accuracy and the inability to do so relying on studies contaminated by the confession-verification confound, only one study to date has tackled directly the confession-bias problem that characterizes field research (Patrick & Iacono, 1991b), and we thus focus on the results of that investigation here. In that RCMP field study involving over 400 cases, we attempted to circumvent the confession-bias confound by reviewing police files for evidence of ground truth that was collected outside of the context of the polygraph examination (e.g., a confession by someone who did not take a polygraph test, a statement that no crime was committed because items believed stolen actually were misplaced). Independent evidence of ground truth was uncovered for one criterion-guilty and 24 criterion-innocent suspects. The fact that it was easier to come by independent evidence of the innocence rather than the guilt of someone taking a CQT stemmed from how the police use polygraph tests to assist their investigations. Polygraph tests typically are administered in cases where the evidence is ambiguous and the police have exhaustively explored available leads to no avail. When a case reaches this point, the investigating officer is hoping that polygraph testing will help resolve the case. Ideally, the suspect will fail and confess, thus giving the investigating officer incriminating evidence that can be used to prosecute the suspect. However, if the suspect merely fails, with no new evidentiary leads to follow, the case is effectively closed, with the police concluding that the individual who failed is guilty. If the suspect passes, the case is often left open, and the search for new suspects and evidence continues.

For those independently confirmed as innocent, the blind rescoring of their polygraph charts produced a hit rate of 57%. Because chance accuracy is 50%, this result indicates the CQT has little better than chance accuracy with the innocent. It also indicates that innocent people are indeed often more disturbed by relevant than control questions. Because only one criterion-guilty person was identified in this investigation, it was not possible to estimate the accuracy of the CQT with persons independently confirmed as guilty.

Despite Patrick and Iacono (1991b) laying out how confession studies bias CQT accuracy estimates and the many subsequent reviews that have echoed this concern about field studies (Fiedler, Schmod, & Stahl, 2002; NRC, 2003), a field study was recently published in a peer-reviewed scientific journal that claimed 100% accuracy for the CQT (Mangan, 2008). This study failed to cite the relevant literature regarding this confession bias problem, and it represents a flawed report that one published commentary characterized as a failure of the peer review system (Verschuere, Meijer, & Merkelbach, 2008; see also Iacono, 2008a).

Q2: Please provide reference entry for Mangan, 2008

Although there are no scientifically credible data regarding the accuracy of the CQT with guilty people, there is reason to doubt the validity of truthful polygraph verdicts. Honts, Raskin, and Kircher (1994) showed that with less than a half hour of instruction regarding CQT theory and how to recognize control and relevant questions, guilty subjects in a mock crime study could learn to escape detection by augmenting their autonomic responses to control questions. They were able to do this using both physical and mental CMs, such as biting the tongue or subtracting 7 serially from a number over 200 when the control question was asked. Moreover, experienced examiners were unable to identify those subjects who employed CMs successfully. The information contained in the instructions given to those escaping detection in this study is widely available in various publications (including in Honts et al., 1994, as well as Lykken, 1998) and on the Web (e.g., www.polygraph.com, https://antipolygraph.org/), making it relatively easy for those so motivated to learn both how the CQT works and how to augment responses to control questions. Subsequent studies by Honts and colleagues (reviewed in Honts & Amato, 2002; see also Honts & Alloway, 2007) have explored how easy it is for naive volunteers to determine on their own how to use CMs and have concluded that uninformed individuals resort to CM strategies that are often ineffective. However, in these studies, the guilty volunteers typically are given little incentive to use CMs effectively, thus leaving their generalizability to real life settings questionable.

DIRECTED LIE TECHNIQUE

Little is known about the validity of the DLT. Although one field study involving the DLT has been published (Honts & Raskin, 1988), this study was also subject to the confession-bias problem. In addition, only a single directed lie question was used, and this question was embedded in a conventional CQT, making it difficult to determine how the test would have fared had directed lie controls been used exclusively. The DLT appears especially susceptible to CMs. When the examiner introduces the directed lies to the subject, they are explained as questions designed to elicit a response pattern indicative of lying. Hence, their purpose is made transparent to subjects, who may understand that an exaggerated response to these questions will help them pass test items on which they lie and presumably offer a less significant response. In addition, the examiner has no idea what issues are covered by the directed lies and how strong an emotional response they are capable of eliciting. For instance, if the subject is directed to answer no to the question “Have you ever done something that you later regretted?” and the subject had an abortion or killed someone in a drunk driving incident, might not the emotions elicited by the directed lie elicit stronger autonomic responses than the material covered by a question concerned with less significant matters, such as theft or fraud?

Guilty Knowledge Test

Of the three classes of polygraph tests considered in this review, only the GKT is spurned by practicing polygraphers. Because of this, few data available from real-life GKT applications can be used to evaluate validity. There are many laboratory simulations of the GKT, and Lykken (1998) has outlined the criteria that define a well-conceived GKT and also reviewed studies that use GKTs meeting these criteria. For instance, Lykken noted that a good test might have 10 items, each with five alternatives, and the person taking the test would be asked to repeat each alternative rather than merely responding no to each, to ensure the examinee was paying attention. The alternatives for each item should be distinctly different from each other, so the examinee can readily recognize the guilty alternative. Lykken's review of eight studies with well-constructed GKTs found accuracy rates of 88% and 97% for guilty and innocent study subjects, respectively.

A meta-analysis of 22 investigations by MacLaren (2001) that used less selective criteria for study inclusion reported somewhat lower accuracies (76% for guilty and 83% for innocent subjects). In a comprehensive meta-analytic review, Ben-Shakar and Elaad (2003) examined 80 studies and included moderator analyses that pointed to several factors that enhanced validity. Studies that employed mock crime simulations, motivational incentives to succeed, verbal responses to item alternatives, and five or more questions produced better hit rates than those without these features. The authors concluded that "the GKT may turn out to be one of the most valid applications [of a test based on] psychological principles" (p. 145). Another study by Ben-Shakar and Elaad (2002) showed that a GKT composed of many questions that focus on numerous aspects of the event at issue has better detection efficiency than a test of identical length that focuses on only one or two aspects of the event. This finding is important because, in field applications, it is often difficult to develop questions, so it is easier to generate a test composed of one or a few items presented repeatedly than a test composed of many different items.

The GKT, as represented in the studies reviewed previously, relies on the measurement of autonomic nervous system measures, most typically the electrodermal response. However, measures of other functions may work as well as or better than autonomic measures. For instance, GKT studies in which brain event-related potentials (ERP) have served as the dependent measure have been similarly impressive in their classification accuracy. Farwell and Donchin (1991) reported perfect classification of "guilty" and "innocent" subjects based on a comparison of their P300 reactions to relevant and irrelevant items of information. A more detailed review of brain-based techniques for assessing deception, including variants of the GKT that have utilized P300 response, is provided below (see the "Alternative Methods" section).

Because the test is virtually never used in North America, no field studies of the GKT have been conducted here. However, the GKT is routinely used in

Japan (Nakayama, 2002), and two studies have been reported by investigators in Israel. Elaad (1990) and Elaad, Ginton, and Jungman (1992) examined the GKT records of 178 criminal suspects tested by examiners from the Israel Police Scientific Interrogation Unit, whose criterion status had been established via confessions. In all but one instance, the GKT was administered following a CQT and included from one to six questions repeated from two to four times, a procedure that, as noted (Ben-Shakhar & Elaad, 2002), diminishes the effectiveness of the GKT. Excluding inconclusive outcomes, innocent examinees were identified with high accuracy (error rate of 2%–3%). Guilty people were less accurately identified, with hit rates varying from 42% to 75% depending on the choice of scoring criteria.

PERSONNEL SCREENING

Because almost everyone recognizes that the RIT is biased against the innocent (e.g., Horowitz, Kircher, Honts, & Raskin, 1997), it has been replaced by the CQT for specific incident investigations. However, despite their lack of empirical foundation, RIT variants and the TES are nevertheless commonly used by the government for employee screening.

Although personnel screening tests that require responses of consistently similar magnitude across many relevant questions to identify truthfulness may appear more credible than the traditional RIT, their premises and applications also have been challenged. Heightened reactions to certain specific questions may occur for reasons other than deceptiveness, such as indignation about being asked the question, exposure to some related issue through the media, or knowledge of someone else who has engaged in the sort of activity covered by that question. Moreover, there is no reason to assume that enhanced reactions to an evocative question will subside once the examinee has offered an explanation for those enhanced reactions to the examiner. In fact, the CQT rests on the opposing (also unproven) assumption that truthful subjects will remain worried about control questions even after these items have been modified to accommodate their admissions. These criticisms give rise to the concern that personnel screening is likely to be associated with a high false positive error rate. In fact, however, as applied by government agents, the false negative error rate seems to be a much more substantial concern, because out of the thousands of personnel screening tests administered every year, only a handful of individuals fail (NRC, 2003).

Research conducted at the NCCA offers some insight into why few individuals fail polygraph screening tests. In their unpublished government report, Barland, Honts, and Barger (1989) described the results of a large analog study that was designed to assess the validity of periodic espionage screening tests administered by experienced government examiners from multiple federal agencies. The 207 study participants were government military and civilian employees. "Guilty" subjects went through complex simulations in which they met with an agent purportedly engaged in espionage who recruited them to collaborate in this activity. Consequent to their recruitment, these "spies" committed acts of mock espionage in which they copied or stole classified documents—just the type of activities that periodic screening tests were designed to detect.

The results of this study indicated a high rate of correct classification for innocent participants (94%) but a low hit rate for guilty participants (34%). The high false negative rate could be related to several factors, but the one most likely is related to the fact that the examiners in this study, who were unaware of the base rate of guilt (about 50%), were following the established field practice of passing almost everyone who took the test. Because periodic screening in real life is in a sense a fishing expedition in which the base rate of spying is presumably negligible, and because examiners are likely to be discouraged from falsely accusing innocent people, many of whom are high-ranking, well-educated, and trained government officials with many years of government service, testing and decision-making practices in the screening context are likely to be biased toward finding few examinees deceptive (Barland et al., 1989; Honts et al., 1994).

Apparently in part because of findings like these, the TES was developed and subjected to two laboratory studies (Department of Defense Polygraph Institute Research Division Staff, 1997, 1998) that reported relatively low rates of both false positive (12.5%) and false negative (17%) error. As noted previously, classification rates observed in analog studies cannot be expected to generalize to the field, where one could expect many innocent government employees with top-secret security clearances to be more bothered by loyalty-challenging questions about espionage and sabotage than directed lie questions about traffic violations. Moreover, because even a 12.5% false positive rate among highly trained weapons lab scientists would wreak havoc on the ability of the United States to carry out its nuclear weapons program, field examiners adjust the threshold for failing the TES so virtually no one fails. The NRC analysis of the TES (NRC, 2003), which included additional unpublished government studies not available to the public, reached the conclusion that "these studies do not provide strong evidence for the validity or utility of polygraph screening" (p. 133).

Q3: Please provide entries for Department of Defense Polygraph Institute Research Division Staff 197 and 1998

NRC's Conclusion Regarding Lie Detection Accuracy

A report from the National Academy of Sciences (NRC, 2003) provides the most comprehensive review of the evidence for polygraph test accuracy ever undertaken. This review, which was requested by the Department of Energy, was launched in part because of concerns regarding the desirability of expanding the government's personnel screening program to include scientists working in the Department of Energy's weapons laboratories. However, the review covered polygraph testing in its entirety, focusing on specific incident polygraph tests because, as we have noted, there are no scientifically peer reviewed, published studies on the validity of screening tests. The review was carried out by a panel of 14 distinguished scientists, with no connection to polygraphy, who represented a variety of disciplines and types of scientific expertise. These scientists had the training, education, and stature to provide a competent and unbiased professional evaluation of the polygraph literature.

Their critique, spread throughout a 398-page volume, was overwhelmingly negative. The panel members did not attempt to estimate precisely polygraph accuracy, nor did they distinguish among types of tests (e.g., CQT versus GKT) or how hit rates may vary for guilty and innocent subjects. Instead, they identified a set of 57 specific incident studies that met "minimal criteria" (NRC, 2003, p. 107) for consideration, noting that the selected studies "do not generally reach the high levels of research quality desired in science" (p. 108). Using the data from these studies, they plotted receiver operating curves (ROCs), borrowing a method from signal detection theory. The primary statistic derived from this analysis was an "accuracy index (*A*)" corresponding to the area under the ROC curve. *A* takes on a value between .5 and 1.00 and, although similar to percentage correct, does not translate directly to the types of percentage estimates reported in the studies analyzed or to those typically reported in reviews of this literature, in part because the ROC analysis takes into account inconclusive outcomes as well as the differences across studies in the rules followed to determine how the outcome of a polygraph test was classified. Because none of the analyzed studies showed the polygraph to have accuracy at or below chance and because these studies indicate well below perfect accuracy, the panel concluded that, for naive examinees untrained in CMs, specific incident polygraph tests have hit rates "well above chance, though well below perfection" (p. 214).

ALTERNATIVE APPROACHES TO DETECTING DECEPTION

A growing area of interest concerns alternatives to conventional polygraph techniques, including reliance on brain ERPs, functional magnetic resonance imaging (fMRI), and thermal imaging techniques. Interest in these methods has been spurred in part by the desire to develop new methods for lie detection that, unlike the CQT, are likely to meet legal standards for what constitutes scientific evidence. This section considers each of these alternative methods in turn.

ERP-BASED DETECTION METHODS

The most extensively researched alternative approach to detection of deception has utilized components of the brain ERP, in particular the P300 component, which occurs in response to significant, infrequent (often referred to as oddball) stimuli. In a P300-based GKT procedure, the crime-relevant keys comprise the rare, meaningful stimuli. When interspersed with the crime-irrelevant multiple-choice alternatives, none of these key alternatives appears "odd" to the person without guilty knowledge, so they elicit minimal P300 response. For the guilty person, the crime-relevant keys are far fewer in number than irrelevant alternatives and are recognized as special, and thus they elicit enhanced P300 reactions.

Q4: Please provide the reference for ''Harrington v. State of Iowa (2001)''.

A real-life example of the use of this approach was in the case of *Harrington v. State of Iowa* (2001). Here, a P300-based GKT was admitted as evidence in the appeal of Terry Harrington, a man who consistently maintained his innocence despite being convicted of murder more than 20 years earlier. Using the procedures outlined in Farwell and Donchin (1991), Harrington was found to have passed a brain ERP-GKT related to his knowledge of the crime scene by showing no enhanced-P300 recognition response to stimuli involving crime details that were identified by Farwell. Moreover, Harrington showed a brain recognition response to stimuli involving his alleged alibi that were developed independently by Farwell without the knowledge or participation of Harrington. Harrington's conviction was ultimately overturned. Farwell has used the term *brain fingerprinting* to refer to this ERP-GKT and formed a company to market its application. Because this methodology, unlike conventional lie detector methods, is based on the strong scientific foundation afforded by decades of research on the GKT (Verschuere, Ben-Shakhar, & Meijer, 2011), we devote special attention to research on the ERP-GKT in this section.

Initial Published Studies. The first published research report of P300 in the detection of guilty knowledge was by Rosenfeld, Nasman, Whalen, Cantwell, and Mazzeri (1987).[1] In this study, participants were shown a box containing nine items (e.g., camera, film, coins), identified the item they would most want to keep, and wrote a 100-word essay describing reasons for this choice. Next, participants viewed a series of words on a monitor, each repeated several times, with instructions to attend carefully to all words. For "guilty" participants ($n = 10$), one of the words (the key) corresponded to the chosen item, with the rest consisting of words for novel items of commensurate value (e.g., radio, cassette, medal). For "innocent" participants ($n = 6$), all of the words consisted of labels for novel items; one of these was arbitrarily designated the key. Statistical analysis of ERP amplitude within a 400- to 700-ms window following word onset revealed significantly larger P300 for the key versus the irrelevant words in the guilty group; statistics were not presented for the innocent group. A practical limitation of this study was that no criteria were presented for classifying individuals as guilty versus innocent. Nevertheless, based on a visual inspection of the waveforms for each individual, the authors concluded that all but one of the guilty participants showed distinct P300 differentiation between key and irrelevant words.

A further limitation of this study was that participants were explicitly instructed to attempt deception by thinking no whenever the key word appeared, which may have contributed to enhanced P300 responses. This feature of the procedure also

[1] A conference abstract summary of a study utilizing a P300-based ERP approach to detection of deception—subsequently reported as Study 2 of an article by Farwell and Donchin (1991)— appeared a year earlier (Farwell & Donchin,1986), at which time the Rosenfeld et al. (1987) report was under editorial review (cf. Rosenfeld, 2011).

Q5: AU: Please provide reference entry for Rosenfeld, 2011 limits external validity, insofar as real-life guilty suspects could not reasonably be expected to comply with such an instruction. Rosenfeld et al. (1988) addressed this issue with a revised protocol in which attention to test words was ensured by instructing participants to look for and count occurrences of one of the novel irrelevant words whenever it appeared on the screen. Results paralleled those of the initial study. Participants in the guilty group ($n = 7$) showed significantly larger P300 responses to the key nontarget word than to irrelevant nontarget words, and for all individuals, responses to the key word exceeded those to irrelevant nontargets (i.e., in no case did amplitude of response to the seven irrelevant nontargets exceed 75% of the amplitude for the key word). A procedural limitation in terms of realism was that participants, as in the study conducted by Rosenfeld et al. (1987), were required to compose an essay regarding the chosen item prior to testing. Other limitations were (a) statistics were not presented for innocent participants ($n = 5$), (b) no quantitative criteria were provided for categorizing participants as guilty versus innocent based on their test responses, and (c) data from three additional guilty participants were excluded from the report due to excessive eye movements or P300 nonresponding.

Two follow-up studies by Rosenfeld and colleagues evaluated the use of ERP measures in more conventional polygraph testing formats. Rosenfeld, Angell, Johnson, and Qian (1991) examined the accuracy of P300 as an index of deception in a procedure analogous to the standard control question test. Rather than testing for knowledge of specific crime details, the test included "Did you do it?" questions pertaining to a specific offense under investigation, along with control questions pertaining to other accusations. Based on a complex, four-step classification algorithm, hit rates for guilty and innocent participants in this study were 92% and 86.6%, respectively. M. M. Johnson and Rosenfeld (1992) evaluated the utility of P300 for detecting deception in a variant of a pre-employment screening test. P300 response was recorded to phrases describing various antisocial acts, presented sequentially on a computer monitor, interspersed with a target phrase to which participants responded with a button press. Upon completion of the test, ground truth was evaluated by having participants complete a checklist under ostensibly anonymous conditions, on which they indicated whether they had committed any of the antisocial acts listed in the ERP test. Hit rates for guilty and innocent participants, based on a three-step classification algorithm, were 100% and 76%, respectively. Although these results appear fairly impressive, the studies themselves are subject to the same sorts of criticisms described earlier with regard to other laboratory investigations of the control question and employee screening tests.

Another influential early article on the use of P300 to detect guilty knowledge was authored by Farwell and Donchin (1991). The two experiments described in this report were innovative in several respects. First, the crime scenarios were quite realistic. In Experiment 1, participants underwent one of two espionage role-plays involving the exchange of information with a "foreign agent," in which they were exposed to six critical details included as probes on the guilty knowledge test.

In Experiment 2, participants were tested about details of minor offenses they had committed in real life. In both experiments, guilt versus innocence was manipulated within subjects (i.e., in Experiment 1, each individual was tested concerning details of the role-play in which he or she participated [guilty condition] as well as the other scenario [innocent condition]; in Experiment 2, each participant was tested regarding the offense he or she had committed [guilty] along with details of another offense committed by a different study participant [innocent]). Another notable feature of these experiments was that the GKT protocol, which paralleled that described by Farwell and Donchin (1986), required participants to respond to all test stimuli: Irrelevant targets (one sixth of trials) prompted a left button press, and irrelevant nontargets (two thirds of trials) and crime-relevant nontargets (probes; one sixth of trials) prompted a right button press. This ensured that participants attended to all stimuli and classified them in a manner that optimized P300 responses.

A further innovation of this study was that it introduced a statistical criterion for classifying participants as innocent or guilty based on comparative P300 responses to irrelevant nontargets and crime-relevant probes. The technique, known as bootstrapping(Efron, 1979), yields an estimate of the sampling distribution for a parameter under circumstances of limited data, by randomly and iteratively sampling from available scores and computing values of the parameter for each subsample. In the Farwell and Donchin (1991study, bootstrapping was used to estimate, for each participant, cross-correlations (i.e., reflecting the degree of relationship between corresponding points of one ERP waveform and another across time) between (a) the average P300 response to probes and the average response to irrelevant non-targets, and (b) the average response to probes and the average to irrelevant targets. If the estimated correlation between probe and target values significantly exceeded that between probe and nontarget values, it was concluded that the participant had recognized the probes as rare and distinctive compared with nontargets and that "guilty knowledge" was present. Conversely, if the correlation between values for probe and nontarget trials exceeded that between probe and target trials, it was concluded that guilty knowledge was not present.

Results were impressive. In Experiment 1, 18 of 20 participants were classified correctly in the guilty condition, with 2 cases inconclusive (i.e., above-mentioned correlations did not differ significantly), and 17 of 20 were correctly classified in the innocent condition, with 3 inconclusives. In Experiment 2, all 4 participants were classified correctly in the guilty condition, and 3 of 4 were correctly classified in the innocent condition, with 1 inconclusive. Thus, in cases for which the bootstrap classification analysis yielded a conclusive outcome, 100% accuracy was achieved.

Nonetheless, there were some notable limitations in this study. Sample sizes were small, particularly in Experiment 2. The accuracy of the test in the guilty conditions was almost certainly enhanced by the fact that, in both experiments, participants explicitly reviewed the crime-relevant details (probe items) prior to taking the test—in contrast to real life, where crime-relevant details are encoded ad hoc and unlikely to be rehearsed prior to testing. Also, no adverse consequences were contingent on test performance, unlike real-life circumstances. Although the presence of threat could augment reactions to critical items among suspects with guilty knowledge, it is also possible that high negative affect might impair memory retrieval and brain response differentiation. A further point is that a simple reaction time (RT) measure (i.e., latency to press the designated button following the stimulus) also differentiated clearly between criterion conditions in Experiment 1: Participants in the guilty condition showed reliably longer RTs to probes versus irrelevant nontargets, whereas in the innocent condition they did not. Although the authors dismissed RT as a viable index of guilt status on the grounds that it can easily be manipulated, findings from subsequent studies have demonstrated that simple CMs can in fact be used to alter P300 responses in a ERP-based GKT (Rosenfeld, Soskins, Bosh, & Ryan, 2004; see “The Impact of Countermeasures,” further on) and that an RT-based GKT actually might be more resistant to CMs (Seymour, Seifert, Shafto, & Mosmann, 2000).

One other early published study that served as a foundation for subsequent work in this area was conducted by Allen, Iacono, and Danielson (1992). Although framed more as a study of memory than deception, this study nonetheless employed a test protocol similar to that of Farwell and Donchin (1991) to assess for the presence of guilty knowledge. Findings were reported for three experiments involving a common protocol. Participants learned two lists of category words, one at the beginning of the experimental session (delayed list) and the other just prior to the P300-based memory test (immediate list), after completing a series of intervening tasks. In the memory test, participants pressed a “yes” button whenever they saw a word from the immediate category list (1/7 of trials), and a different “no” button whenever they saw a word from either the delayed list (1/7 of trials), or from 1 of 5 nonlearned category lists (5/7 of trials). Thus, on the test, participants had to inhibit a tendency to respond to previously learned words in the same way as words they had just learned. Recognition of words from the delayed list was predicted to yield enhanced P300 response in comparison with nonlearned words.

A key feature of this study was that it relied on a novel statistical technique for classifying individual participants as knowledgeable or not with regard to specific word lists, a Bayesian classification strategy. This involved selecting various parameters of the ERP waveform that differentiated learned from unlearned words (e.g., P300 amplitude; area under the curve within 200 ms on either side of the P300 peak) and then using information about the discriminability of these parameters and the relative frequencies of learned and unlearned trials to compute a probability for each participant that an ERP average for a given word list reflected one or the other condition. The discrimination parameters and Bayesian classification algorithm were developed using data from 20 participants in Experiment 1 and then cross-validated on two new samples of 20 participants each in Experiments 2 and 3. Procedures were identical across experiments except that (a) instructions differed slightly in Experiments 2 and 3 (i.e., participants were told to press yes for words they had learned and no for words they had not, but to deliberately lie about words from the initial learned list by pressing the "no" button to these words, and (b) participants in Experiment 3 were promised $5 if they could control their brain responses so as to prevent detection of words they had lied to.

Using the Bayesian algorithm and cutpoints developed in Experiment 1, the sensitivities (probability of correctly classifying a learned list as learned; cf. true positive rate) in Experiment 2 and 3 were .925 and .95, respectively, compared with .95 in Experiment 1. The specificities (probability of correctly classifying an unlearned list as unlearned; cf. true negative rate) were .94 and .98, respectively, compared with .96 in Experiment 1. In a reanalysis of data from this study, Allen and Iacono (1997) found that the use of Farwell and Donchin's (1991) bootstrapping method to classify lists as learned versus unlearned yielded no incorrect classifications; however, it yielded inconclusive results for learned lists in 13% of cases and for unlearned lists in 28% of cases. Allen et al. 1992 also examined the accuracy of classifications based on two indicators of behavioral response to words from each list (RT, response errors); sensitivities were .95 and .95 in Experiments 2 and 3, respectively, and specificities were .95 and .98 (versus .975 and 1.0, respectively, in Experiment 1). Thus, classification accuracies based on behavioral response indices were commensurate with those based on ERP parameters (cf. Farwell & Donchin, 1991). They were also in line with the findings of Seymour et al. (2000), who found that RTs to probe stimuli could be used to separate guilty from innocent individuals in a RT-based GKT, even when subjects were instructed to modify their responses to escape detection.

These results indicated that concealed information could be detected with very high accuracy in individual cases using a probabilistic analysis of ERP response parameters. However, some limitations of the Allen et al. (1992) study are important to consider in relation to detection of deception in real-life cases. In particular, the word-learning task has limited external validity vis-à-vis a real-world crime situation. Simple category words are obviously very different from crime-relevant details. Also, as in other work cited, participants in this study explicitly learned the relevant words as opposed to encountering them incidentally in a dynamic real-world context. Furthermore, the Bayesian classification algorithm developed in Experiment 1 capitalized on information that may not readily be available in real-life cases—namely, the ground-truth status of previously learned lists. ERP parameters were selected in part because they discriminated words on these "concealed" lists from words on the unlearned lists. With real-world suspects, the status of information as concealed or not is normally indeterminate. Although a parallel algorithm could be developed using ERP data from real-life cases in which a solid ground-truth criterion (e.g., a corroborated confession; DNA evidence) became available after testing, the generalizability of this algorithm to cases different from those included in the development sample (e.g., in terms of type of crime, latency since commission, suspect characteristics, etc.) would be open to question. With regard to these points, it should be reiterated that the Allen et al. study was framed as an investigation of memory rather than of deception. Nevertheless, issues such as these are important to consider in applying the findings of this study to the problem of detecting deception.

Subsequent Studies Building on Initial Published Work. The most active researcher in this area over the past two decades in terms of published studies in peer-reviewed journals has been Peter Rosenfeld of Northwestern University. Many of the studies reported by Rosenfeld and his colleagues through the early 2000s, following the approach of Allen et al. (1992), focused on P300 as an index of dissimulated ("malingered") amnesia for simple types of learned material, such as words, numbers, and basic autobiographical facts, rather than details of an enacted "crime" (for reviews of this work, see Rosenfeld, 2002; Rosenfeld & Ellwanger, 1999). Other investigations of this type were published during this period by Allen and colleagues (Allen, Iacono, Laravuso, & Dunn, 1995; Allen & Movius, 2000; van Hoof, Brunia, & Allen, 1996; for reviews of this work, see Allen, 2002; Allen & Iacono, 2001).

Building on this basic work investigating P300-based detection of generic learned information, research over the past decade has focused on further evaluating the effectiveness of ERP methods for detecting crime-relevant knowledge in investigative contexts. One series of studies by Rosenfeld and colleagues, on the impact of CMs on detectability using P300-based methods, is discussed in the next subsection. Another line of work, by Lawrence Farwell and colleagues, has focused on a scoring method termed MERMER (memory and encoding related multifaceted electroencephalographic response) that entails quantification of multiple features of the ERP response to test stimuli, including the P300 along with other parameters. In an initial full-length report of this quantification method by Farwell and Smith (2001), six participants were tested, three of them regarding known biographical details from their own lives and the other two regarding unfamiliar biographical details. The test protocol, like that of Farwell and Donchin (1991), was a response task that included irrelevant target stimuli (calling for a left button press) along with irrelevant nontargets and crime-relevant nontargets (each calling for a right button press). Hit rates for both conditions in this study ("guilty"-informed, "innocent"-uninformed) were reported as 3/3 (100%). Subsequent studies of this method have evaluated its accuracy in mock crime (Farwell, Hernandez, & Richardson, 2006) and actual or simulated field contexts (Farwell, 2008; Farwell et al., 2006; Farwell, Richardson, & Richardson, 2011), but these studies have been reported only in conference abstract form. In a recent review of studies using the MERMER scoring approach, Farwell (2012) characterized this technique as yielding 100% accurate classifications in all research studies to date, with no "indeterminate" (inconclusive) outcomes. The lack of indeterminate outcomes was cited as an advantage of the MERMER scoring approach over the more standard P300-focused scoring approach.

These reported findings for the MERMER method have been criticized on several grounds. As noted by Rosenfeld (2005), a serious limitation of this work from the standpoint of scientific evaluation is that the quantification parameters for MERMER are not described in sufficient detail in any published report to permit replication, because they are patented and considered proprietary. Although Farwell and Smith (2001) stated that MERMER scoring entails quantification of the parietal P300 and a subsequent negative-polarity component, maximal at frontal sites, along with "phasic changes in the frequency and structure of the [ERP] signal" (p. 137), the nature of these latter "phasic changes" was not specified in this article or in subsequent reports by Farwell and colleagues. Farwell's (2012) review article does clarify that the bootstrap cross-correlation approach of Farwell and Donchin (1991) serves as the basis in MERMER for evaluating similarity of ERP components across differing stimulus conditions but fails to specify how (a) evaluations for P300 and late-negative components are combined, or (b) phasic signal changes are incorporated into the waveform morphology comparisons.

Rosenfeld (2005) also raised other concerns regarding the MERMER scoring technique. He questioned Farwell et al.'s characterization of the MERMER approach as yielding 100% accuracy of classifications based on only a single published journal article, when studies published by other investigative groups using P300-based approaches had reported accuracies below this level. In addition, Rosenfeld challenged the scientific status of the additional late-negative component and "phasic change" parameters utilized in the MERMER approach. Whereas an extensive literature shows that P300 is sensitive to the salience/recognizability of presented information, the functional significance of the other MERMER parameters is unclear. Citing Soskins, Rosenfeld, and Niendam (2001), Rosenfeld (2005) pointed out that, although the late-negative component in part reflects recovery to baseline of the preceding P300 response, a parameter that may contribute incrementally to detectability of known versus unknown information, it also may contain some nonspecific, artifact-related variance (i.e., associated with capacitive rebound of the signal at filter settings used for recording of P3).

Regarding the "phasic change" parameters used in MERMER, Rosenfeld (2005) pointed out: "The meaning of these other claimed independent (but undocumented) frequency phenomena, which, according to Farwell himself, are not found in all persons, is another matter. . . . *The supportive data—e.g., power spectra illustrating these claimed frequency effects—have never been shown anywhere*" (p. 27 [emphasis in original]). Some of these concerns raised by Rosenfeld were echoed in a more recent critique of the MERMER technique by Meijer, Ben-Shakar, Verschuere, and Donchin (2012). Additionally, these authors challenged the assertion made by Farwell (2012) that his brain fingerprinting method had been evaluated in studies involving "over 200 test cases"; they pointed out that the set of studies cited by Farwell overlapped substantially in terms of participant samples and included many unpublished datasets, such that peer-reviewed findings pertaining to the method's validity are in fact limited to results from a total of only 30 participants across two peer-reviewed journal articles (Farwell & Donchin, 1991; Farwell & Smith, 2001).

With regard to evaluation of ERP-based methods in applied contexts, a further study that warrants mention is one by Mertens and Allen (2008), which employed a virtual reality crime procedure to evaluate the accuracy of a P300-based detection of deception test. Participants in the study logged on to a computer in an unoccupied office and navigated through a highly realistic virtual environment depicting the interior of a multi-room apartment. Innocent participants were instructed simply to explore the virtual apartment for a designated period of time. Guilty participants entered the virtual apartment for a similar period of time under instructions to "steal" specified items from the apartment through use of a computer mouse. Following exposure to the virtual environment, participants underwent a P300-based detection test akin to that of Farwell and Donchin (1991), including probe (crime-relevant), target (learned irrelevant), and distracter (nonlearned irrelevant) items. Guilty participants completed the detection test either without instruction regarding how to defeat the test (subgroup 1 = no CMs) or under instructions to perform specific types of CMs (subgroup 2 = mental CMs to target stimuli; subgroup 3 = physical CMs to target stimuli; subgroup 4 = alternating physical and mental CMs to distracter stimuli).

An additional feature of the study was that classification accuracy was compared for three different scoring methods: bootstrapped cross-correlation (Farwell & Donchin, 1991), bootstrapped peak-to-peak amplitude difference (Rosenfeld et al., 2004; Soskins et al., 2001), and Bayesian probability analysis (Allen et al., 1992). For innocent participants, the cross-correlation method produced a very high rate of indeterminate outcomes (56%), with the remainder of cases (44%) correctly classified. By contrast, the two other scoring methods yielded conclusive classifications for all innocent participants, with accuracy for the peak-to-peak method (100%) slightly exceeding that for the Bayesian method (96%). In the case of uninstructed (non-CM) guilty participants, the indeterminate rate for the cross-correlation method was again very high (60%), with 27% of cases correctly classified as guilty and 13% incorrectly classified as innocent. For these same guilty participants, the peak-to-peak and Bayesian methods each produced 47% correct ("guilty") classifications and 53% incorrect ("innocent") classifications, with no indeterminate outcomes. Results for the guilty CM groups are discussed in the next subsection.

Based on these results, Mertens and Allen (2008) concluded that the accuracy of P300-based detection tests with guilty suspects may be appreciably lower in field contexts involving memory for real-life crime details as compared to lab contexts involving learned lists of probe items. At the same time, these investigators noted that the P300-based detection method—in contrast with the conventional control question procedure used by North American polygraph examiners—is advantageous in terms of yielding very low rates of false positives (i.e., innocent cases mistakenly classified as guilty). In sum, these authors concluded that guilty/deceptive outcomes of ERP-based detection tests are likely to be of substantial value for investigative decision making in real-life cases (i.e., because such outcomes are strongly diagnostic of the presence of crime-relevant knowledge), whereas innocent/truthful outcomes are likely to be less accurate and thus of questionable utility.

Impact of Countermeasures on P300-Based Information Detection. A major development over the past few years has been the emergence of systematic research on the effects of CMs on the accuracy of ERP-based detection tests and on approaches to detecting and reducing their effectiveness. Whereas an initial study by Sasaki, Hira, and Matsuda (2001) focusing on use of a simple mental CM manipulation (counting backward by sevens) reported no impact on P300-based detection, a subsequent article by Rosenfeld et al. (2004) reported significant effects on P300 responsivity and detectability as a function of physical and affective-imaginal CMs. This article reported findings from two experiments testing for guilty knowledge, the first entailing crime scenarios akin to those used in Experiment 1 of Farwell and Donchin (1991) and the other a simpler procedure entailing detection of personal knowledge (i.e., participant's birth date).

Because procedures aside from the "crime" manipulation (including scoring methods) were similar, only the first experiment is detailed here. In Rosenfeld et al.'s (2004) Experiment 1, guilty participants ($n = 22$) participated in one of two mock crime scenarios, one involving the theft of a ring from a desk drawer and the other involving the removal of an official university grade roster from its posted location. Innocent participants ($n = 11$) did not participate in either scenario. To ensure encoding of crime details, guilty participants rehearsed the instructions for their scenario several times and were required to pass a test pertaining to the critical details before undergoing the ERP-based GKT procedure. The test stimuli were a series of single words from various categories, some representing crime-relevant details (probes; 17% of trials) and others irrelevant category words, with some of these designated targets (17% of trials) and the remainder nontargets (66% of trials). As in Farwell and Donchin (1991), participants were instructed to press one button whenever a target word appeared and a different button whenever a nontarget word (whether crime-relevant or irrelevant) appeared. Half of the guilty participants received training in CM use prior to the GKT, and the others did not. CMs entailed specific covert responses (e.g., pressing a finger against a leg; wiggling a big toe; imagining being slapped) following the appearance of irrelevant nontarget words in order to enhance brain responses to these stimuli.

Two bootstrap techniques were used to classify participants as guilty or innocent on the basis of ERP responses. One of these involved comparing the estimated correlation between probe and target response with that between probe and nontarget response (cf. Farwell & Donchin, 1991). The other entailed comparing the difference in mean amplitude of response to probes versus irrelevant nontargets for each individual participant against a distribution of difference scores formed by iterative resampling of the available data. Within each analysis, participants were classified as innocent if a significant difference in the direction indicating guilt was not obtained (differences were evaluated in terms of both base-to-peak and peak-to-peak amplitude; results for the more effective, peak-to-peak score analysis are described here).

Using the correlation-difference method, 10 of 11 innocent participants (90.9%) were correctly classified, but only 6 of 11 in each of the simple-guilty and guilty-CMs groups (54.5%) were correctly classified. For the amplitude-difference method, 10 of 11 innocent participants (90.9%) and 8 of 11 simple-guilty participants (72.7%) were correctly classified, but only 2 of 11 guilty-CMs participants (18.2%) were correctly classified. In Experiment 2, hit rates for guilty-CMs participants were: correlation difference method, 3/12 (25%); amplitude-difference method, 6/12 (50%). Comparative hit rates without CMs for these same participants were 62.9% and 92.3% when tested prior to instruction in and use of CMs and 25% and 58.3% when tested again after instruction/use of CMs.

Some interpretive difficulties are evident in this study. No inconclusive category was employed in classifications, making it difficult to compare these findings with those of Farwell and Donchin (1991). In addition, the hit rate in Rosenfeld et al.'s (2004) Experiment 1 for simple-guilty participants based on the correlation-difference method (6/11 = 55%) was substantially and inexplicably lower than the rate for guilty participants in Experiment 1 of the Farwell and Donchin study; even with inconclusives considered as incorrect, the hit rate across the two experiments in this earlier study was 22/24 = 91.7%. This comparatively unimpressive hit rate for non−CM participants in this experiment clouds interpretation of the low hit rate for CM participants. Interpretation of CM effects in Experiment 2 was likewise complicated by differences in non−CM hit rates across for two separate comparison sessions as well as by the artificiality and narrowness of the test procedure (i.e., focus on a single biographic detail). Notwithstanding these limitations, the Rosenthal et al. study was important in raising concerns about deliberate CMs being used to beat an ERP-based detection test and inspiring further research on this topic.

The next published investigation of the impact of CMs was conducted by Mertens and Allen (2008), whose results from the no-CM guilty condition were summarized in the preceding section. This study was notable for its highly realistic virtual reality theft scenario that served as the crime manipulation, inclusion of multiple CM conditions (mental CMs to target stimuli, physical CMs to target stimuli, and alternating physical and mental CMs to distracter stimuli), and comparison of differing approaches to the scoring of test data. Whereas the hit rate for non-CM guilty participants based on the optimal method of scoring (either peak-to-peak or Bayesian) was 47%, the maximum hit rate for any of the CM conditions achieved by any method of scoring was only 27%.

A third, bootstrapped cross-correlation scoring method produced very high rates of indeterminate decisions (56%−93%) in all study conditions, including the innocent condition. As with the Rosenfeld et al. (2004) study, the modest detection rate for non-CM guilty participants reported by Mertens and Allen (2008; which, by implication, casts further doubt on the 100% across-the-board accuracy rate for MERMER-based detection claimed by Farwell [2012]) complicates interpretation of the low detection rate reported for CM participants. Nonetheless, the hit rates for CM groups in this study were significantly lower than rates for the non-CM group, corroborating Rosenfeld et al.'s (2004) conclusion that CMs can be effective in reducing the accuracy of ERP-based detection.

In response to this emerging evidence for the effectiveness of CMs, Rosenfeld et al. (2008) sought to develop an alternative, CM-resistant ERP test protocol. In this procedure, termed the complex trial protocol (CTP), two stimuli are presented in sequence on each test trial, separated by a varying interstimulus interval of ~1 to 2 seconds. The first (S1) consists of either a rare probe stimulus (20% probability), relevant to the matter under investigation, or a frequent irrelevant stimulus (80%). To this initial stimulus, the participant responds with a standard designated button press, regardless of stimulus type, to signify registration of the S1. The second stimulus (S2) appears within 1.2—1.8 seconds after offset of the first and comprises either a target or non-target stimulus calling for differential button press responses. The intent of the CTP procedure is to separate the processing of relevant and irrelevant test stimuli from the discriminative response task required to generate a referent against which to compare responses to test stimuli, in the form of target P3 response. Although delayed, the target (S2) part of the task serves to maintain attention and ensure compliance with the task.

Rosenfeld et al. (2008) reported this task to be highly accurate in the detection of personal (birth date) information, yielding correct decisions for 12 of 12 non-CM participants in two separate studies when an individualized ("flexible") search window was used to define P300 and correct decisions for 11 of 12 participants trained to use the same types of CMs employed by Rosenfeld et al. (2004). These authors found that the use of CMs led to enhancement of P300 responses to both probe and irrelevant S1 stimuli, which they attributed to increased attentional processing required to decide whether to initiate CMs upon presentation of each stimulus. To test the hypothesis that the observed P300 differentiation between probes and targets for CM participants in this task might be attributable in part to omission of CMs selectively for probe stimuli, Meixner and Rosenfeld (2010) instructed participants to selectively omit CMs for one of five S1 stimuli in a CTP while employing CMs with the others and demonstrated the occurrence of an enhanced P300 response to the omit-CM stimulus relative to the employ-CM stimuli. Further, they found that, if the omit-CM stimulus was personally meaningful in some way, the degree of response augmentation for this stimulus was markedly larger. In subsequent work, Rosenfeld and Labkovsky (2010) showed that, although use of CMs for some but not all irrelevant stimuli in a CTP (i.e., for two of four irrelevants, as opposed to four of four) resulted in elimination of this omission-enhancement effect, the detection rate for CM guilty participants under these instructional conditions was nonetheless very high (100%).

A potential weakness of the CTP (noted by Farwell, 2012) is that participants are not required to differentiate behaviorally between probe and irrelevant stimuli presented as S1s (i.e., a common button press response is made in each case). Consequently, it is conceivable that participants motivated to defeat the test could avoid processing S1 stimuli in the test beyond the level of detecting visible changes in the foreground display associated with their occurrence and responding accordingly (i.e., without registering the distinct features of individual stimuli).

A contrasting perspective on this issue comes from work by Lui and Rosenfeld (2009) demonstrating enhancement of P300 response to a dishonestly answered probe stimulus when preceded by a subliminal presentation of the stimulus (i.e., very brief occurrence, in the context of a surrounding visual mask). The authors reported an overall accuracy rate of 86% for this method across guilty and innocent participants in this study. Although this study did not examine the impact of CMs on detectability using this method, it will be interesting in future work to examine whether a subliminal priming manipulation can be incorporated into a CTP procedure in a manner designed to protect against deliberate inattentiveness to S1 stimuli. More broadly, it will be important to evaluate the effectiveness of methods such as CTP and subliminal priming in more highly realistic experimental scenarios such as that used by Mertens and Allen (2008).

Other ERP Response Components. In addition to P300, other components of brain potential response have been applied to the detection of deception. One is the N400 response that reliably occurs in response to semantic incongruity (i.e., words that complete a sentence in an unexpected fashion; Kutas & Hillyard, 1980). Boaz, Perry, Raney, Fischler, and Shuman (1991) developed an N400-based GKT procedure in which participants, after viewing either a tape of an enacted burglary or a noncrime control tape, were presented with crime-relevant phrases that concluded with either true or false endings. Hit rates in this study (73.2% overall in cross-validation samples) were markedly lower than in most P300-based GKT studies to date.

Subsequently, Fang, Liu, and Shen (2003) explored the use of contingent negative variation (CNV) in detection of deception. The CNV is a slow negative shift in electroencephalogram potential that develops during anticipation of a target stimulus following presentation of a warning cue. Fang et al. examined CNV in a task in which participants were first presented with face stimuli and then, upon presentation of a follow-up signal, indicated whether the face was familiar to them or not. These authors reported significantly enhanced CNV on trials in which participants prepared to enact a false response compared with trials on which they responded truthfully. The comparative promise of this method for detecting deception is difficult to evaluate, because no effort was made to classify participants as truthful or deceptive on the basis of brain response. Further work with this type of procedure is needed to evaluate its effectiveness at the level of individuals.

Two other trends in the use of brain response measures to detect deception are noteworthy. One consists of studies designed to link differing components of the ERP to specific cognitive *processes* underlying deception (cf. Furedy, Davis, & Gurevich, 1988). Along this line, R. Johnson, Barnhardt, and Zhu (2003) reported evidence for two distinct components of the ERP connected with the act of deception, one reflecting inhibition of the prepotent (truthful) response and the other reflecting monitoring of past truthful and deceptive responses (see also R. Johnson, Barnhardt, & Zhu, 2004). Subsequently, R. Johnson, Barnhardt, and Zhu (2005) reported differential effects of practice (i.e., trial repetitions) on reaction time and ERP parameters of response to test questions—indicating that cognitively mediated response conflicts underlying deceptive behaviors are resistant to practice effect, in a manner similar to perceptually driven response conflicts.

In contrast with the ERP studies just reviewed, the focus of work of this kind is on gaining insights into the dynamics of neurocognitive processing associated with deception rather than on classifying individuals as truthful or deceptive based on ERP parameters that discriminate these conditions empirically. Notably, this has been a prominent focus to date in neuroimaging studies of deception, reviewed in the next subsection.

Neuroimaging-Based Detection Methods

A major development over the past decade has been the growing use of neuroimaging measurement in research on deception. Studies of this kind have utilized the technique of fMRI, in which changes in blood flow within specific regions of the brain are indexed by perturbations in a magnetic field surrounding the head, or in some cases (e.g., Abe, Suzuki, Mori, Itoh, & Fujii, 2007; Abe et al., 2006) positron emission tomography (PET), an imaging technique in which neural activity in specific brain regions is indexed through measurement of subatomic particles emitted by a radioactive isotope injected into the brain.

As in the ERP work of Johnson and colleagues (Johnson, 2003, 2004), many studies of this type have focused on processes associated with deception (and affiliated brain regions) rather than on classifying participants as deceptive or truthful. The first such study was by Spence et al. (2001), which reported enhanced activation in the ventrolateral prefrontal cortex (Brodmann area 47) bilaterally when participants lied about activities they had performed earlier in the day. This activation was interpreted as reflecting an inhibitory process associated with the effort to withhold the truth. Two subsequent studies reported increased activity in a wider array of brain regions (including frontal/prefrontal, parietal, and temporal cortices) when participants lied to critical items on a GKT (Langleben et al., 2002) or a GKT-like memory test (Lee et al., 2002).

In another early study (Ganis, Kosslyn, Stose, Thompson, & Yurgelun-Todd, 2003), researchers examined activations associated with two distinct parameters of a lie: (1) whether it is spontaneous or rehearsed and (2) whether it is isolated or part of a broader story the participant is telling. Well-rehearsed lies connected to a broader narrative evoked greater activation in right anterior frontal cortex than spontaneous isolated lies, whereas spontaneous isolated lies elicited greater activation in anterior cingulate and posterior visual cortices. Lies of both types evoked greater activation (versus truth-telling) in right and left anterior prefrontal cortex and parahippocampal gyrus, right precuneus, and left cerebellum. These findings indicate that different brain regions are recruited in the context different forms of lying activity.

In the decade or so since publication of these initial studies, the neuroimaging literature on deception has expanded rapidly. A pervasive finding has been increased activity in regions of the prefrontal cortex—including bilateral dorsolateral and anterior regions (middle and superior frontal gyri) and inferior frontal regions—during deception or concealment of information (Abe, 2009). Other brain regions that have been implicated in deception include the angular gyrus, caudate nucleus, and supplementary motor area. A major question arising from this work has to do with the specificity of the role that these brain regions play in deception. Prominent investigators in this area (e.g., Kozel, Padgett, & George, 2004; Langleben et al., 2005) have argued that brain activations indexed by neuroimaging are more revealing of the basic cognitive processes underlying deceptive responding than peripheral response measures, which index more generic bodily activation. Alternatively, it is conceivable that activations reliably reported in neuroimaging studies of deception reflect engagement of brain systems that play a supportive role in many contexts calling for concentrated attention and cognitive control—as opposed to systems that mediate deception or information concealment per se (cf. Gamer, 2011).

In addition to studies focusing on identifying brain regions associated with the act of deception, a number of studies have examined the effectiveness of fMRI-based assessment for classifying individuals as deceptive versus truthful. An initial study along this line by Kozel et al. (2004) examined the consistency with which particular brain regions were activated across participants during lying as compared to truth telling. Some degree of consistency was evident, in anterior brain regions in particular, encouraging further work.

In a follow-up study involving more than 60 participants, Kozel et al. (2005) evaluated the accuracy of fMRI-based testing for classifying individuals as truthful or deceptive in relation to the commission of a mock theft. Participants stole a watch or a ring under instruction and then underwent a test protocol that resembled a control question polygraph test sequence. The test included crime-relevant questions, neutral questions dealing with facts and personal preferences, and control questions dealing with illegal or rule-breaking behaviors of differing types. Data for half the sample were used to identify patterns of brain activation that differentiated deception from truthful responding. This resulted in three anatomic clusters (centered around the right orbitofrontal/inferior frontal cortex, right middle frontal gyrus, and anterior cingulate cortex) being selected as discriminating. These clusters were then used as regions of interest for classifying participants in the remainder of the sample.

More specifically, the number of activated voxels for each of these areas in the deceptive versus truthful condition was calculated for each participant in the second half-sample with reference to an a priori statistical threshold. The resulting difference score was then used to classify each participant as deceptive or truthful with respect to one or the other mock theft based on whether the difference was significant in a positive or a negative direction. Based on this approach, 28 of 31 participants (93%) were classified correctly with respect to the theft they had committed (watch or ring). Corresponding rates in three subsequent replication samples (Kozel, Johnson et al., 2009; Kozel, Laken et al., 2009) were 71%, 93%, and 86%, for an average rate of 85.8% across all four cross-validation samples, including that of Kozel et al. (2005).

Findings such as these have engendered considerable enthusiasm around the possibilities of "direct" brain-based detection of deception. Alongside the growing cadre of studies in this area, commercial enterprises have surfaced in the United States that offer neuroimaging-based detection of deception services (e.g., Cephos Corporation, www.cephoscorp.com; No Lie MRI, Inc., www.noliemri.com). Stephen Laken, who was a coauthor on the Kozel reports and founder of Cephos, unsuccessfully testified in court to have exculpatory fMRI results admitted in *United States v. Semrau* (2012; see Shen & Jones, 2011, for analysis of this decision). In *State of Maryland v. Gary Smith* (2012, see Shen & Jones, 2011), scientists associated with No Lie MRI attempted to achieve the same outcome for the defendant in state court but were similarly unsuccessful. Considering that most of the available published research on the use of fMRI-based testing for classifying individuals as deceptive versus truthful has utilized basic "Did you do it?" question formats analogous to conventional RIT or CQT test procedures, this trend toward the use of a discredited questioning format is cause for some concern. As noted, it is quite conceivable that the brain activations that differentiate deceptive and truthful conditions in lab studies are indicative of more general cognitive processes such as focused attention, working memory, and cognitive control. It is also unclear to what extent activation of similar regions might occur in innocent individuals responding to "Did you do it?" questions under the conditions of uncertainty and anxiousness that tend to characterize real-life detection tests. The one individual-classification study that included a no-crime innocent group (Kozel, Johnson, et al., 2009) reported an accuracy rate of only 8 out of 21 (38%) for a CQT-type test in this group. This result casts doubt on the effectiveness of standard detection test protocols with innocent participants, even when based on functional neuroimaging methodology. Others have also expressed considerable skepticism regarding the evidentiary value of fMRI findings in court (Bizzi et al., 2009; Greeley & Illes, 2007; Wagner, 2010).

Q6: Please provide the reference for "State of Maryland v. Gary Smith 2012".

Some more recent evidence, however, does indicate that neuroimaging-based detection can achieve higher rates of classification of innocent subjects through used of a GKT-like (concealed information test) format. The first study to examine brain activations associated with deception in a GKT-type test was one by Gamer, Bauermann, Stoeter, and Vossel (2007). This study focused on average condition effects rather than classification of individuals. However, a subsequent study by Nose, Murai, and Taira (2009) that focused on classification of individual participants in a GKT-type test procedure reported accuracy rates of 84% for both guilty/deceptive and innocent/truthful participants. Further research is needed to evaluate whether this more costly and technically demanding approach to GKT testing carries any advantage relative to ERP-based or more conventional autonomic-response based testing.

THERMAL IMAGING

Thermal imaging has also been used to detect deception by employing a high-speed motion picture camera sensitive to rapid changes in facial regional blood flow. For example, in a mock crime study by Pavlidis, Eberhardt, and Levine (2002), 6 of 8 guilty and 11 of 12 innocent subjects were correctly identified based on an undescribed "thermal signature," apparently involving changes in blood flow around the eyes, in relation to an incident involving theft of $20.

This method is intriguing, because it may be possible to use it without the subject's knowledge, potentially under conditions of remote testing (e.g., through a computer–video interface). A high-profile example of an application of this kind came to the attention of the public in 2011 when it was announced that thermal imaging technology would be tested in an undisclosed U.K. airport as a security screening device. That same year, an empirical study was published that evaluated the accuracy of thermal imaging for this specific purpose (Warmelink et al., 2011). In this study, 51 passengers in an international airport either lied or told the truth about a forthcoming trip in an on-site interview conducted by study experimenters. Skin temperature was recorded using a thermal imaging camera. On the basis of increases in facial skin temperature during the interview, 69% of deceptive participants and 64% of truthful participants were classified correctly. The authors noted that judgments of veracity made by interviewers after interacting with participants achieved higher rates of accuracy (77% and 72%, respectively) than the thermal imaging–based classifications. The authors concluded that thermal imaging is unlikely to have incremental validity over standard questioning methods for purposes of airport screening.

More research is needed to evaluate this technique in other contexts and to determine whether it might be vulnerable to many of the same criticisms leveled at conventional polygraph tests. In particular, it seems likely that, in real-life testing situations, a considerable portion of falsely accused innocent people would show heightened facial blood flow when asked a threatening question they answer honestly. The results of the Warmelink et al. (2011) study, which yielded a hit rate of only 64% for innocent participants, appear consistent with this possibility.

Voice Stress Analysis

One technique unlikely to be of any value in the detection of deception is voice stress analysis. Recent heightened concerns about security have led to an increase in interest in this technique, which involves analyzing a sample of human speech for effects presumed due to inaudible microtremors of the vocal muscles reflective of the stress of lying. The advantage of voice stress analysis is that it can be used with recorded or broadcast speech without the subject's knowledge. The disadvantage is that, despite 30 years of research, there is virtually no evidence for its scientific basis or that it accurately detects lying (NRC, 2003).

THE POLYGRAPH IN COURT

Two important considerations in courtroom presentation of polygraph findings are the admission of polygraph testimony into evidence and how juries evaluate this evidence.

Admission of Polygraph Testimony

Polygraph tests often find their way into criminal court through one of two routes. One involves the stipulated test in which polygraph examinations are administered with the prior agreement of prosecutor and defense attorney. Often the prosecution will agree to a stipulated test when the case against the defendant is weak. In these circumstances, if the suspect passes the test, the charges are dismissed. If the suspect fails the test, the prosecution reserves the right to submit the polygraph findings to the court. About half of U.S. states endorse the use of stipulated tests, but Canadian courts refuse them.

Another way that polygraph results may enter a courtroom is over the objection of the prosecution in cases where it is argued that polygraph results constitute valid scientific evidence. This practice is allowed by law in New Mexico (*Lee v. Martinez*, 2004), provided the polygraph test administration satisfies certain standards. It is also a strategy increasingly adopted by defense attorneys who wish to determine if current circumstances favor the admission of a polygraph test that they have arranged for a client who subsequently passed. Often a hearing is requested before a judge who is asked to determine if polygraph tests satisfy standards for scientific evidence in light of new laws and rulings and/or in light of recent developments in the field (e.g., computerization) that may indicate polygraphy has been improved significantly since the last time the court considered admitting such evidence.

In 1923, in *Frye v. United States*, the U.S. Supreme Court established the rules for determining the admissibility of testimony based on novel scientific techniques in federal proceedings In this case, James Frye was denied the opportunity to have considered as evidence the results of a polygraph test administered by psychologist William Marston, the "father" of modern polygraphy. Although the *Frye* ruling no longer controls federal proceedings, it is still influential to the laws of some states that followed the *Frye* precedent of requiring "general acceptance" of a technique by the relevant scientific community before testimony based on the technique can be admitted. In federal courts as well as in many state courts, the standards that control are those laid out by the U.S. Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals* (1993). These standards direct that judges are to consider the admissibility of testimony based on newly developed scientific procedures after consideration of a number of factors including, but not limited to, whether the procedure (a) is supported by scientific theory, (b) has been subjected to peer review, (c) has a known error rate, (d) is governed by uniform standards, and (e) is generally accepted in the relevant scientific community.

Hence, following motions submitted by defense attorneys, many courts hold hearings based on principles outlined in *Frye* or *Daubert* to determine if testimony based on a defendant's passed polygraph test should be admitted as evidence (see Faigman, Saks, Sanders, & Cheng, 2009, for a more thorough review of the legal status of polygraph testing in the United States). Such hearings are likely to be influenced by the Supreme Court's decision in *United States v. Scheffer* (1998), in which it ruled that defendants in military court martial proceedings do not have a right to admit as evidence the results of exculpatory polygraph tests, based on the justices' ruling that there is no consensus in the scientific community that polygraph evidence is valid.

When a defense attorney arranges for a client to take a polygraph test, the results of the test are protected by attorney–client privilege. If the defendant fails the test, the results would not be divulged, because doing so would only serve to undermine the defendant's credibility. A test administered under these circumstances is considered to be "friendly." Such a test stands in contrast to an "adversarial" test administered by the law enforcement personnel, the results of which would be made known to the prosecution and defense. Because fear of the consequences of being detected is considered to be important to the valid outcome of a test, and there appears to be less to lose and therefore less to fear with a friendly test, it seems likely that friendly tests would be easier to pass than adversarial tests. Moreover, because the defendant is paying the polygrapher with the hope of passing the test, the examiner is being pressured, at least by the defendant, to produce the desired outcome. In a procedure that is as subjective and unstandardized as the CQT, it is easy to imagine how subtle adjustments to the procedure could increase the likelihood of friendly tests being passed. Unfortunately, there are no empirical studies attesting to the validity of friendly tests. All the existing field studies deal with adversarial tests.

HOW JURIES EVALUATE POLYGRAPH EVIDENCE

An important issue surrounding the use of polygraph evidence in court is the weight that is likely to be attached to this evidence by juries. This concern derives in part from Rule 403 of the Federal Rules of Evidence (and its state court equivalents, see Daniels, 2002), which allows courts to exclude evidence if its probative value is substantially outweighed by the prejudicial impact it may have on the jury. Unlike other types of evidence a jury may hear, polygraph evidence has the potential to usurp the jury's constitutionally mandated task of deciding guilt. Thus, courts have also excluded polygraph testimony on the grounds that the scientific and technical aura that surrounds the practice of polygraph testing may lead juries to assign excessive probative weight to this evidence (see, e.g., *United States v. Alexander*, 1975).

Since our review of how juries consider polygraph evidence in the last edition of this text, one new study has appeared (Myers, Latter, & Abdollahi-Arena, 2006). This study and those that have preceded it suggest that mock juries are skeptical of polygraph test results. However, just as field studies are needed to estimate the accuracy of polygraph tests, it would be worthwhile to have data from polled jurors following trials in which polygraph testimony was offered to determine how jurors weighed this evidence in actual legal proceedings.

SCIENTIFIC OPINION

The opinions of scientists regarding polygraphy are obviously important. Conventional polygraph tests have a weak conceptual foundation. Moreover, serious methodological problems that are unlikely to be easily overcome make it unlikely that any line of research will yield findings that resolve concerns about accuracy. Given this state of affairs, there is considerable value in the broad-based sampling of the opinions of scientists with the background and expertise to evaluate polygraph tests. In addition, the *Frye* and *Daubert* decisions make clear that the views of the scientific community about the general acceptance of a technique are important to considering admissibility of testimony based on the technique.

Only one investigation of scientific opinion regarding polygraph techniques has been published in a scientific peer-reviewed journal (Iacono & Lykken, 1997). This study polled members of the Society for Psychophysiological Research and fellows in Division 1 (General Psychology) of the American Psychological Association (APA). High response rates (>74%) were obtained from those in both organizations, and there was remarkable agreement across groups regarding CQT polygraphy. These scientists expressed a high level of skepticism regarding the claims of polygraph proponents. They did not find the theory of the CQT to be scientifically sound or the accuracy claims of polygraph proponents to be credible. In addition, they expressed opinions indicating that friendly tests have little value and CMs pose a significant threat to the validity of passed tests. Members of neither group would recommend that testimony based on the results of CQTs be admitted in court.

Only APA members were asked about directed lie tests, and they did not agree that these tests are scientifically sound. In contrast to these negative opinions about conventional specific incident tests, those polled had favorable opinions about the GKT. The contrast in the scientific credibility of the CQT and GKT is important, because it indicates that respondents were not generally skeptical about detection of deception techniques but have doubts that are specific to the CQT. The results of these surveys parallel the opinions of the NRC (2003) committee that reviewed polygraph test validity as well as those of many other scientists and professional societies at arm's length from the polygraph profession have conducted critical appraisals of the field (APA, 2004; Ben-Shakar, 2002; British Psychological Association, 2004; Fiedler et al., 2002; Oksol & O'Donohue, 2003; Verschuere et al., 2008; Vrij, 2008).

CONCLUSION

Despite this scientific skepticism, the use of polygraph tests continues unabated, presumably reflecting beliefs among law enforcement and national security policy makers that their utility benefits outweigh concerns regarding costs associated with their misuse. There appears to be little dispute about the utility of polygraph testing, although only anecdotal, not scientific evidence, exists to support this contention (NRC, 2003). Nevertheless, many criminal suspects confess following failed tests, providing a means to resolve criminal investigations that otherwise would go unprosecuted.

In employee screening, the admissions employees make about their alcohol use, sex lives, and colleagues' suspect behavior provide the government with what is considered to be valuable information that would be virtually impossible to obtain via any other (legal) means. Likewise, those administering sex offender treatment programs have come to rely on polygraph tests to encourage offenders to divulge fully their past sexual misdeeds, so much so that the use of polygraph tests in these programs is now widespread. When used in such contexts, the polygraph is little more than a prop intended to encourage socially undesirable self-disclosure among those who believe it genuinely works, a phenomenon established over 40 years ago as the "bogus pipeline" effect (Jones & Sigall, 1971). However, as the NRC panel noted, in the long run, evidence that a technique lacks validity will eventually undercut its utility.

For many decades, polygraph testing has been part of the fabric of our institutions for law enforcement and national security. Consequently, reliance on polygraphy as an investigative tool is unlikely to diminish in the future. Although it remains possible that the CQT will become accepted as credible scientific evidence, courts have not shown a readiness to embrace the admission of specific incident tests in the first 15 years following *Daubert* (Faigman et al., 2009). As our review indicates, there is little evidence to support their admission, and what evidence does exist, coupled with the obvious weaknesses in CQT theory, indicates that the CQT has little more than chance accuracy with innocent people and can be easily defeated by guilty people who learn to augment their responses to control questions.

Although the GKT and the ERP-GKT appear to offer promising alternatives to the CQT (Ben-Shakhar, 2012; Ben-Shakhar, Bar-Hillel, & Kremnitzer, 2002, Iacono, 2011), research with these procedures has not focused on how to adapt them successfully to field applications (Iacono, 2008b, 2011). fMRI and other methods that are not based on the GKT have produced a body of research that is vulnerable to the same criticisms that have been leveled against the CQT. As scientists who have worked in this area for over 30 years, we are struck by the fact that this literature has focused on pushing the technological prowess of fMRI while neglecting the importance of the strong research designs that a half century of CQT research has taught us are needed to credibly anchor validity claims for lie detection methods.

Despite a lack of adequate field study and standardized test protocols, the GKT is based on sound theory, and it is possible for a jury to weigh evidence regarding the adequacy of a GKT. Consider, for instance, how much weight might be assigned a properly conducted GKT indicating the presence of guilty knowledge. Assume suspect John Fisbee is asked to preapprove the questions on a 12-item GKT by indicating whether he knows the answer to any of the questions, and he claims no knowledge. In addition, after the test is administered, he is asked if he can guess the answers to any of the items, and the two items he "guesses" the correct answer to are eliminated from further consideration. The test is given by an examiner who is unaware of the correct answers. On the GKT, Fisbee shows the strongest physiological response to all of the guilty alternatives for the remaining questions. When the same test was given to 10 individuals, none of whom could be involved in the crime, they responded to the guilty alternatives at chance levels. Because it is difficult to understand how such an outcome could come about in the absence of Fisbee's guilty knowledge, such a test result provides relatively strong prima facie evidence of guilt. One can alter aspects of this hypothetical scenario in various ways (e.g., Fisbee fails 8 of the remaining 10 items), but with each alteration, it is possible to make a scientifically informed appraisal regarding the level of confidence one can have in the outcome. By contrast, passing a GKT is much more difficult to interpret because the field research needed to determine what those who commit crimes are likely to remember has not been conducted. Until this work is carried out, passed GKTs will remain suspect.

REFERENCES

Abe, N. (2009). The neurobiology of deception: Evidence from neuroimaging and loss-of-function studies. *Current Opinion in Neurology*, *22*, 594–600.

Abe, N., Suzuki, M., Mori, E., Itoh, M., & Fujii, T. (2007). Deceiving others: Distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. *Journal of Cognitive Neuroscience*, *19*, 287–295.

Abe, N., Suzuki, M., Tsukiura, T., Mori, E., Yamaguchi, K., & Itoh, M. (2006). Dissociable roles of prefrontal and anterior cingulate cortices in deception. *Cerebral Cortex*, *16*, 192–199.

Allen, J. J. B. (2002). The role of psychophysiology in clinical assessment: ERPs in the evaluation of memory. *Psychophysiology*, *39*, 261–280.

Allen, J. J. B., & Iacono, W. G. (1997). A comparison of methods for the analysis of event-related potentials in deception detection. *Psychophysiology*, *34*, 234–240.

Allen, J. J. B., & Iacono, W. G. (2001). Assessing the validity of amnesia in dissociative identity disorder: A dilemma for the DSM and the courts. *Psychology, Public Policy, and Law*, *7*(2), 311–344.

Allen, J. J. B., Iacono, W. G., & Danielson, K. D. (1992). The identification of concealed memories using the event-related potential and implicit behavioral measures: A methodology for prediction in the face of individual differences. *Psychophysiology*, *29*, 504–522.

Allen, J. J. B., & Movius, H. L. (2000). The objective assessment of amnesia in dissociative identity disorder using event-related potentials. *International Journal of Psychophysiology*, *38*, 21–41.

Allen, J. J., Iacono, W. G., Laravuso, J. J., & Dunn, L. A. (1995). An event-related potential investigation of posthynoptic recognition amnesia. *Journal of Abnormal Psychology*, *104*(3), 421–430.

American Psychological Association. (2004, August 5). The truth about lie detectors (aka polygraph tests). Retrieved from http://www.apa.org/research/action/polygraph.aspx

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

Ben-Shakar, G. (2002). A critical review of the control questions test (CQT). In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 103–126). San Diego, CA: Academic Press.

Ben-Shakhar, G. (2012). Current research and potential applications of the concealed information test: an overview. *Frontiers in Psychology*, *3*, 342. doi:10.3389/fpsyg.2012.00342

Ben-Shakhar, G., Bar-Hillel, M., & Kremnitzer, M. (2002). Trial by polygraph: reconsidering the use of the guilty knowledge technique in court. *Law and Human Behavior*, *26*(5), 527–541.

Ben-Shakhar, G., & Elaad, E. (2002). Effects of questions' repetition and variation on the efficiency of the guilty knowledge test: A reexamination. *Journal of Applied Psychology*, *87*(5), 972–977.

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *Journal of Applied Psychology*, *88*, 131–151.

Bizzi, E., Hyman, S. E., Raichle, M. E., Kanwisher, N., Phelps, E. A., Morse, S. J., . . . Greely, H. T. (2009). *Using imaging to identify deceit: Scientific and ethical questions*. Cambridge, MA: American Academy of Arts & Sciences.

Boaz, T. L., Perry, N. W., Raney, G., Fieschler, I. S., & Shuman, D. (1991). Detection of guilty knowledge with event related potentials. *Journal of Applied Psychology*, *76*, 788–795.

British Psychological Association. (2004, October 6). *A review of the current scientific status and fields of application of polygraphic deception detection*. Leicester, England: Author.

Daniels, C. W. (2002). Legal aspects of polygraph admissibility in the United States. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 327–338). San Diego, CA: Academic Press.

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Elaad, E. (1990). Detection of guilty knowledge in real-life criminal investigation. *Journal of Applied Psychology*, *75*, 521–529.

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology*, *77*, 757–767.

Employee Polygraph Protection Act, Pub. L. 100–347 (1988).

Faigman, D. L., Saks, M. J., Sanders, J., & Cheng, E. K. (2009). The legal relevance of scientific research on polygraph tests. In D. L. Faigman, M. J. Saks, J. Sanders, & E. K. Cheng (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 5, pp. 266–297). Eagan, MN: Thomson Reuters/West.

Fang, F., Liu, Y., & Shen, Z. (2003). Lie detection with contingent negative variation. *International Journal of Psychophysiology*, *50*, 247–255.

Farwell, L. A. (2008). Brain fingerprinting detects real crimes in the field despite one-hundred-thousand-dollar reward for beating it. *Psychophysiology*, *45*, S104.

Farwell, L. A. (2012). Brain fingerprinting: A comprehensive tutorial review of detection of concealed information with event-related brain potentials. *Cognitive Neurodynamics*, *6*, 115–154.

Farwell, L. A., & Donchin, E. (1986). The "Brain Detector": P300 in the detection of deception. *SPR Abstracts*, *23*, 434.

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event related brain potentials. *Psychophysiology*, *28*, 531–547.

Farwell, L. A., Hernandez, R. S., & Richardson, D. C. (2006). Brain fingerprinting in laboratory conditions. *Psychophysiology*, *43*, S38.

Farwell, L. A., Richardson, D. C., & Hernandez, R. S. (2006b). Brain fingerprinting in field conditions. *Psychophysiology*, *43*, S38.

Farwell, L. A., Richardson, D. C., & Richardson, G. (2011). Brain fingerprinting field studies comparing P300-MERMER and P300 ERPs in the detection of concealed information. *Psychophysiology*, *48*, S95–S96.

Farwell, L. A., & Smith, S. S. (2001). Using brain MERMER testing to detect knowledge despite efforts to conceal. *Journal of Forensic Sciences*, *46*, 1–9.

Ferguson, R. J. (1966). *The polygraph in private industry*. Springfield, IL: Charles C Thomas.

Fiedler, K., Schmod, J., & Stahl, T. (2002). What is the current truth about polygraph lie detection? *Basic & Applied Social Psychology*, *24*, 313–324.

Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).

Furedy, J. J., Davis, C., & Gurevich, M. (1988). Differentiation of deception as a psychological process: A psychophysiological approach. *Psychophysiology*, *25*, 683–688.

Gamer, M. (2011). Detecting of deception and concealed information using neuroimaging techniques. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Theory and application of the Concealed Information Test* (pp. 90–113). New York, NY: Cambridge University Press.

Gamer, M., Bauermann, T., Stoeter, P., & Vossel, G. (2007). Covariations among fMRI, skin conductance, and behavioral data during processing of concealed information. *Human Brain Mapping*, *28*, 1287–1301.

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: An fMRI investigation. *Cerebral Cortex*, *13*, 830–836.

Ganis, G., & Patnaik, P. (2009). Detecting concealed knowledge using a novel attentional blink paradigm. *Applied Psychophysiology & Biofeedback*, *34*(3), 189–196. doi: 10.1007/s10484-009-9094-1

Greeley, H. T., & Illes, J. (2007). Neuroscience-based lie detection: The urgent need for regulation. *American Journal of Law & Medicine*, *33*, 377–431.

Harrington v. State of Iowa, 109 F. 3d 1275; Court of Appeals, 8th Cir. (1997).

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

Honts, C. R., & Alloway, W. R. (2007). Information does not affect the validity of a comparison question test. *Legal & Criminal Psychology*, *12*, 311–320.

Honts, C. R., & Amato, S. (2002). Countermeasures. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 251–264). London, England: Academic Press.

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

Honts, C.R., Raskin, D., & Kircher, J. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology*, *79*, 252–259.

Honts, C. R., Raskin, D., & Kircher, J. (2002). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 2, pp. 446–483). St. Paul, MN: West.

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

Horvath, F. (1977). The effect of selected variables on the interpretation of polygraph records. *Journal of Applied Psychology*, *62*, 127–136.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), *Advances in psychophysiology* (pp. 201–207). London, England: Jessica Kingsley.

Iacono, W. G. (2007). Polygraph testing. In E. Borgida & S. T. Fiske (Eds.), *Beyond common sense: Psychological science in the courtroom* (pp. 219–235). Oxon, England: Blackwell.

Iacono, W. G. (2008a). Accuracy of polygraph techniques: Problems using confessions to determine ground truth. *Physiology & Behavior*, *95*, 24–26.

Iacono, W. G. (2008b). Effective policing: Understanding how polygraph tests work and are used. *Criminal Justice & Behavior*, *35*, 1295–1308.

Iacono, W. G. (2008c). The forensic application of "brain fingerprinting": Why scientists should encourage the use of P300 memory detection methods. *American Journal of Bioethics*, *8*, 30–32.

Iacono, W. G. (2010). Lie detection. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology* (Vol. 2, pp. 928–930). Hoboken, NJ: Wiley.

Iacono, W. G. (2011). Encourging the use of the guilty knowledge test (GKT): What the GKT has to offer law enforcement. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 12–23). New York, NY: Cambridge University Press.

Iacono, W. G., & Lykken, D. T. (1997). The validity of the lie detector: Two surveys of scientific opinion. *Journal of Applied Psychology*, *82*, 426–433.

Iacono, W. G., & Lykken, D. T. (2009). The case against polygraph tests. In D. L. Faigman, M. J. Saks, J. Sanders, & E. K. Cheng (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 5, pp. 342–406). Eagan, MN: Thomson Reuters/West.

Iacono, W. G., & Patrick, C. J. (1987). What psychologists should know about lie detection. In I. B. Weiner & A. K. Hess (Eds.), *The handbook of forensic psychology* (pp. 460–489). New York, NY: Wiley.

Iacono, W. G., & Patrick, C. J. (1999). Polygraph ("lie detector") testing: The state of the art. In A. K. Hess & I. B. Weiner (Eds.), *The handbook of forensic psychology* (2nd ed., pp. 440–473). New York, NY: Wiley.

Iacono, W. G., & Patrick, C. J. (2006). Polygraph ("lie detector") testing: Current status and emerging trends. In I. B. Weiner & A. K. Hess (Eds.), *The handbook of forensic psychology* (3rd ed., pp. 552–588). Hoboken, NJ: Wiley.

Johnson, M. M., & Rosenfeld, J. P. (1992). Oddball-evoked P300-based method of deception detection in the laboratory II: Utilization of non-selective activation of relevant knowledge. *International Journal of Psychophysiology*, *12*, 289–306.

Johnson, R., Barnhardt, J., & Zhu, J. (2003). The deceptive response: Effects of response conflict and strategic monitoring on the late positive component and episodic memory-related brain activity. *Biological Psychology*, *64*, 217–253.

Johnson, R., Barnhardt, J., & Zhu, J. (2004). The contribution of executive processes to deceptive responding. *Neuropsychologia*, *42*, 878–901.

Johnson, R., Barnhardt, J., & Zhu, J. (2005). Differential effects of practice on the executive processes used for truthful and deceptive responses: An event-related brain potential study. *Cognitive Brain Research*, *24*, 386–404.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*, 349–364.

Kosson, D. S. (1988). Psychopathy and dual-task performance under focusing conditions. *Journal of Abnormal Psychology*, *105*, 391–400.

Kozel, F. A., Johnson, K. A., Grenesko, E. L., Laken, S. J., Kose, S., & Lu, X. (2009). Functional MRI detection of deception after committing a mock sabotage crime. *Journal of Forensic Sciences*, *54*, 220–231.

Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biological Psychiatry*, *58*, 605–613.

Kozel, F. A., Laken, S. J., Johnson, K. A., Boren, B., Mapes, K. S., & Morgan, P. S. (2009). Replication of functional MRI detection of deception. *Open Forensic Science Journal*, *2*, 6–11.

Kozel, F. A., Padgett, T. M., & George, M. S. (2004). A replication study of the neural correlates of deception. *Behavioral Neuroscience*, *118*, 852–856.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205.

Langleben, D. D., Loughead, J. W., Bilker, W. B., Ruparel, K., Childress, A. R., Busch, S. I., & Gur, R. C. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. *Human Brain Mapping*, *26*, 262–272.

Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., . . . Childress, A. R. (2002). Brain activity during simulated deception: An event-related functional magnetic resonance study. *Neuroimage*, *15*, 727–732.

Lee, T. M., Liu, H. L., Tan, L. H., Chan, C. C., Mahankali, S., Feng, C. M., . . . Gao, J. H. (2002). Lie detection by functional magnetic resonance imaging. *Human Brain Mapping*, *15*, 157–164.

Lee v. Martinez, Supreme Court of New Mexico, 96 P.3d 291 (2004).

Lui, M., & Rosenfeld, J. P. (2009). The application of subliminal priming in lie detection: Scenario for identification of members of a terrorist ring. *Psychophysiology*, *46*, 889–903.

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

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*, 258–262.

Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist*, *29*, 725–739.

Lykken, D. T. (1998). *A tremor in the blood: Uses and abuses of the lie detector* (2nd ed.). New York, NY: Plenum Press.

–MacLaren, V. V. (2001). A quantitative review of the guilty knowledge test. *Journal of Applied Psychology*, *86*(4), 674–683.

Meijer, E. H., Ben-Shakar, G., Verschuere, B., & Donchin, E. (2013). A comment on Farwell (2012): Brain fingerprinting: A comprehensive tutorial review of detection of concealed information with event-related brain potentials. *Cognitive Neurodynamics*, *7*, 155–158.

Meijer, E. H., Verschuere, B., Merckelbach, H. L., & Crombez, G. (2008). Sex offender management using the polygraph: a critical review. *International Journal of Law and Psychiatry*, *31*(5), 423–429.

Meixner, J. B., & Rosenfeld, J. P. (2010). Countermeasure mechanisms in a P300-based concealed information test. *Psychophysiology*, *47*, 57–65.

Mertens, R., & Allen, J.J.B. (2008). The role of psychophysiology in forensic assessments: Deception detection. *Psychophysiology*, *45*, 286–298.

Myers, B., Latter, R., & Abdollahi-Arena, M. K. (2006). The court of public opinion: Lay perceptions of polygraph testing. *Law and Human Behavior*, *30*(4), 509–523. doi:10.1007/s10979-006-9041-0

Nakayama, M. (2002). Practical use of the concealed information test from criminal investigation is Japan. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 49–86). San Diego, CA: Academic Press.

National Defense Authorization Act, Pub. L. 106–65 (2000).

National Research Council. (2003). *The polygraph and lie detection*. Washington, DC: National Academies Press.

Nose, I., Murai, J., & Taira, M. (2009). Disclosing concealed information on the basis of cortical activations. *Neuroimage*, *44*, 1380–1386.

Oksol, E. M., & O'Donohue, W. T. (2003). A critical analysis of the polygraph. In W. T. O'Donohue & E. R. Levensky (Eds.), *Handbook of forensic psychology: Resource for mental health and legal professionals* (pp. 601–634). San Diego, CA: Academic Press.

Patrick, C. J., & Iacono, W. G. (1991a). A comparison of field and laboratory polygraphs in the detection of deception. *Psychophysiology*, *28*(6), 632–638.

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

Pavlidis, I., Eberhardt, N. L., & Levine, J. A. (2002). Seeing through the face of deception: Thermal imaging offers a promising hands-off approach to mass security screening. *Nature*, *415*, 435.

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

Research Division Staff, Department of Defense Polygraph Institute (1997). A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope polygraph and the test for espionage and sabotage. *Polygraph*, *26*, 79–106.

Research Division Staff, Department of Defense Polygraph Institute (1998). Psychophysiological detection of deception accuracy rates using the test for espionage and sabotage. *Polygraph*, *27*, 68–73.

Rosenfeld, J. P. (2002). Event-related potentials in the detection of deception, malingering, and false memories. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 265–286). New York, NY: Academic Press.

Rosenfeld, J. P. (2005). "Brain fingerprinting": A critical analysis. *Scientific Review of Mental Health Practice*, *4*, 20–37.

Rosenfeld, J. P., Angell, A., Johnson, M., & Qian, J. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average waveforms. *Psychophysiology*, *38*, 319–335.

Rosenfeld, J. P., Cantwell, B., Nasman, V. T., Wodjdac, V., Ivanov, S., & Mazzeri, L. (1988). A modified, event-related potential-based guilty knowledge test. *International Journal of Neuroscience*, *42*, 157–161.

Rosenfeld, J. P., & Ellwanger, J. W. (1999). Cognitive psychophysiology in detection of malingered cognitive deficit. In J. J. Sweet (Ed.), *Forensic neuropsychology: Fundamentals and practice* (pp. 287–312). Lisse, the Netherlands: Swets & Zeitlinger.

Rosenfeld, J. P., & Labkovsky, E. (2010). New P300-based protocol to detect concealed information: Resistance to mental countermeasures against only half the irrelevant stimuli and a possible ERP indicator of countermeasures. *Psychophysiology*, *47*, 1002–1010.

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., & Chedid, E. (2008). The Complex Trial Protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology*, *45*, 906–919.

Rosenfeld, J. P., Nasman, V. T., Whalen, R., Cantwell, B., & Mazzeri, L. (1987). Late vertex positivity in event-related potentials as a guilty knowledge indicator: A new method of lie detection. *Polygraph*, *16*(4), 258–263.

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, *41*, 205–219.

Rule 403 Federal Rules of Evidence, Pub. L. 93–595, §1, January 2, 1975, 88 Stat. 1932; April 26, 2011, effective December 1, 2011.

Samuels, D. J. (1983). What if the lie detector lies? *Nation*, *237*, 566–567.

Sasaki, M., Hira, S., & Matsuda, T. (2001). Effects of a mental countermeasure on the physiological detection of deception using the event-related brain potentials. *Japanese Journal of Psychology*, *72*, 322–328.

Seymour, T. L., & Fraynt, B. R. (2009). Time and encoding effects in the concealed knowledge test. *Applied Psychophysiology and Biofeedback*, *34*(3), 177–187. doi:10.1007/s10484-009-9092-3

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess "guilty knowledge" *Journal of Applied Psychology*, *85*, 30–37.

Shen, F. X., & Jones, O. D. (2011). Brain scans as evidence: Truths, proofs, lies, and lessons. *Mercer Law Review*, *62*. 861–883.

Soskins, M., Rosenfeld, J. P., & Niendam, T. (2001). Peak-to-peak measurement of P300 recorded at 0.3 Hz high pass filter settings in intraindividual diagnosis: Complex vs. simple paradigms. *International Journal of Psychophysiology*, *40*, 173–180.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *NeuroReport*, *12*, 2433–2438.

Thurber, S. (1981). CPI variables in relation to the polygraph performance of police officer candidates. *Journal of Social Psychology*, *113*, 145–146.

United States v. Alexander, 526 F. 2d 161. 168 (1975 (8th Cir.)).

United States v. Scheffer, WL 141151 U.S. (1998).

United States v. Semrau, No. 1:07-cr-10074-1. United States Court of Appeals, 6th Circuit (2012)

Van Hoof, J. C., Brunia, C. H. M., & Allen, C. J. (1996). Event-related potentials as indirect measures of recognition memory. *International Journal of Psychophysiology*, *21*, 15–31.

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (2011). *Memory detection: Theory and application of the concealed information test*. Cambridge, England: Cambridge University Press.

Verschuere, B., Crombez, G., De Clercq, A., & Koster, E. H. (2005). Psychopathic traits and autonomic responding to concealed information in a prison sample. *Psychophysiology*, *42*(2), 239–245. doi:10.1111/j.1469-8986.2005.00279.x

Verschuere, B., Meijer, E., & Merkelbach, H. (2008). The quadri-track zone comparison technique: It's just not science. A critique to Mangan, Armitage, and Adams (2008). *Physiology & Behavior*, *95*, 27–28.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). West Sussex, England: Wiley.

Wagner, A. D. (2010). Can neuroscience identify lies? *A judge's guide to neuroscience: A concise introduction* (pp. 13–25). Santa Barbara: University of California.

Waid, W. M., Orne, M. T., & Wilson, S. K. (1979). Effects of level of socialization on electrodermal detection of deception. *Psychophysiology*, *16*, 15- 22.

Warmelink, L., Vrij, A., Mann, S., Leal, S., Forrester, D., & Fisher, R. P. (2011). Thermal imaging as a lie detection tool at airports. *Law and Human Behavior*, *35*, 40–48.