# Scientific Basis for Polygraph Testing

**Raymond Nelson**


## Abstract

Published scientific literature is reviewed for comparison question polygraph testing and its application to diagnostic and screening contexts. The review summarizes the literature for all aspects of the testing procedure including the pretest interview, test data collection, test data analysis, and a proffer of the physiological and psychological basis for polygraph testing. Polygraph accuracy information is summarized for diagnostic and screening exams. Evidence is reviewed for threats to polygraph accuracy and the contribution of polygraph results to incremental validity or increased decision accuracy by professional consumers of polygraph test results. The polygraph is described as a probabilistic and non-deterministic test, involving both physiological recording and statistical methods. Probabilistic tests, statistical models, and scientific tests in general are needed when neither deterministic observation nor physical measurement are possible. Event-specific diagnostic polygraphs have been shown to provide mean accuracy of .89 with a 95% confidence range from .83 to .95. Multi-issue screening polygraphs have been shown to provide accuracy rates, with a mean of .85 and a 95% confidence range of .77 to .93.


Keywords: *Polygraph, lie detection, signal detection, test data analysis, scientific basis.*

Polygraph examinations, like other scientific and forensic tests, can take the form of either diagnostic test or screening tests. The difference between diagnostic and screening exams is that diagnostic examinations involve the existence of a known problem, in the form of symptoms, evidence, allegations, or incidental circumstances that suggest an individual may have some involvement, for which the examination results are intended to support a positive or negative diagnostic conclusion. Screening tests include all tests conducted in the absence of a known incident, known allegation, or known problem.

The purpose of diagnostic tests is to form a conclusion that may serve as a basis for action. This action will often affect the future of an individual in term of rights, liberties or health. For this reason, it is difficult to imagine an ethical justification for the selection of a testing technique that provides something less than the highest achievable level of diagnostic accuracy. Diagnostic tests achieve high levels of decision accuracy, in part, by restricting to the test to a single issue of concern.

In contrast, screening tests are intended to add incremental validity to risk management decisions that are made in the absence of any known problem. This is accomplished both by gathering information and by investigating the possible involvement of an individual in one or more issues of concern. Screening tests should not be used alone as the basis for action that may affect an individual´s rights, liberties, or health. Absence of any known problem is the defining characteristic of a screening test (Wilson, & Jungner, 1968; Raffle, & Muir Gray, 2007). Screening polygraphs tests address the objective of adding incrementally

Author note: Raymond Nelson is a polygraph examiner and psychotherapist who has authored numerous papers and studies on many aspects of the polygraph test. Mr. Nelson is a researcher associated with the Lafayette Instrument Company, is the curriculum director of an accredited polygraph training program, and is an elected member of the Board of Directors of the American Polygraph Association - currently serving in the role of President. Mr. Nelson has taught at several polygraph training programs, is an active speaker and presenter at both international conferences, and has been an expert witness in several court jurisdictions regarding both polygraph and psychotherapy matters. The views and opinions expressed in this publication are those of the author and not necessarily those of the Lafayette Instrument Company or the American Polygraph Association. Inquiries and correspondence can be sent to raymond.nelson@gmail.com.

to risk management through a combination of smaller goals that may include both the discriminate ability of the test result, and the capability of the testing process to develop information. Polygraph screening programs can also involve a third goal in the form of increased deterrence of problems (American Polygraph Association, 2009a; 2009b). Deterrence objectives may be achieved by deterring higher-risk persons from access to high-risk environments (e.g., police applicant screening or government/operational security screening), or through dissuasion of (or decreased) non-compliance with policies, rules, and regulations (e.g., operational security policies).

## Discussion

According to the American Polygraph Association (2011), polygraph examinations consist of three phases: 1) a pretest interview, 2) an in-test data collection phase, and 3) test data analysis. Each of these phases has an important effect on both test accuracy and the usefulness of the test result. For this reason, all assumptions and procedures considered fundamental to the polygraph test should ideally be based on generally accepted knowledge or evidence and theoretical constructs for which there exists published and replicated empirical support.

### Polygraph pretest interview

At its most basic, an interview is merely a conversation with a purpose (Hodgson, 1987), and, as indicated by Kahn and Cannel (1957), the success of many professional endeavors depends in part on the ability to get information from others. The polygraph pretest interview is intended to orient the examinee to the testing procedures, the purpose of the test, and the investigation target questions. The basic premise of interviewing holds that people will report more useful information when they are prompted to do so by an interested listener who builds rapport through the use of conversation and interview questions. Polygraph pretest interviews are intended to allow truthful examinees to become accustomed to - or habituated to - the cognitive and emotional impact of hearing and responding to test stimulus questions that describe their possible involvement in problematic behaviors, while also sensitizing or increasing the awareness and response potential of deceptive examinees to test questions that describe their past behavior.

The polygraph pretest interview is a process, involving several steps (American Polygraph Association, 2009a, 2009b; Department of Defense Polygraph Institute, 2002), including: a free-narrative interview (Powell & Snow, 2007), semi-structured interview (Lindlof & Taylor, 2002), or structured interview (Drever, 1995), a thorough review of the test question stimuli, and a practice or orientation test. The first objective of the pretest interview is to establish a positive identification and introduction, and to clarify the roles of the examiner and examinee. The examiner will also introduce the examinee to the examination room, including the use of audio or video recording devices, and all of the polygraph sensors that will later be attached to the examinee.

The next stage of the process consists of making an initial determination of the suitability of the examinee and obtaining informed consent for testing. This is done after a review of the rights of the examinee during testing, including the right to terminate the examination at any time. Ideally, informed consent should also include information about who will receive the information and results from the examination, and where to obtain more information about the strengths and limitations of the polygraph procedure. The examiner will then engage the examinee in a brief discussion about the case background and personal background of the examinee, in order to continue to establish an adequate and suitable testing rapport. The examiner will also provide more information about the psychological and physiological basis for the polygraph test, and will provide answers to any questions the examinee may have regarding the testing procedures.

A practice test or acquaintance test should be conducted as part of standardized field practice (American Polygraph Association, 2009a, 2009b; Department of Defense, 2006a). The purpose of this test is to orient the examinee to the testing procedure before commencing the actual examination. Research by Kircher, Packard, Bell & Bernhardt (2001) has shown that this can contribute to increased test accuracy. A scientific view, supported by recent studies involving non-naive examinees who are fully aware of the details of the testing procedures (Honts & Reavy, 2009; Honts & Alloway, 2007; Nelson, Handler, Blalock & Hernandez, 2012; Rovner, 1986), holds that the effectiveness of evidence-based scientific tests is not dependent on the examinee's belief system. The purpose of the acquaintance test is not to demonstrate or convince the examinee to believe that the polygraph test is infallible, but to orient the examinee to the testing procedures. Regardless of the examiner's and examinee's attitudes or beliefs concerning the acquaintance test, scientific studies (Bradley & Janisse, 1981; Horneman & O'Gorman, 1985; Horowitz, Kircher & Raskin, 1986; Kirby, 1981; Widup, R, Jr & Barland, 1994) have shown that the use of an acquaintance test does not harm and may at times increase the accuracy of the polygraph examination result. The actual reason for this effect may have more to do with ensuring that the instrument and sensors are adjusted and functioning adequately and that the examinee has had an opportunity to practice complying with behavioral instructions.

The next stage of the pretest interview will be a free-narrative interview, a structured interview or semi-structured interview. Free-narrative interviews are characterized by the use of simple and common language, an absence of coercive techniques, an opportunity for the interviewee to communicate details at the level of one's own choosing, along with encouragement to elaborate. Free narrative interviews conducted during polygraph testing may include direct or probing questions regarding a known or alleged incident, before proceeding to construct polygraph test questions. Free-narrative interview strategies are useful during diagnostic investigations, but are not well suited toward use in polygraph screening tests which are conducted in the absence of a known or alleged incident. Pretest interviews for screening exams conducted during polygraph screening exams, whether pertaining to operational security, law enforcement pre-employment, or post-conviction supervision, will take the form of either a structured interview or semi-structured interview.

Structured interviews differ from semi-structured interview in that structured interviews are conducted verbatim, without deviation from the interview protocol (General Accounting Office, 1991; Campion, Campion, & & Hudson, 1994; Kvale, 1996). In contrast, semi-structured interviews are conducted using a structured content and question outline, for which the interviewer is permitted to present interview questions in a manner that is individualized based on the personalities, education levels, and rapport between the interviewer and interviewee. Although structured interviews may be preferred by some researchers and program administrators for their consistency, structured interviews make little use of the skill and expertise of the interviewer.

Semi-structured interviews are intended to make more effective use of interviewer skill and expertise to access rich information regarding the interview content. Like structured interviews, semi-structured interviews should be anchored by a defined interview schedule or interview protocol, with clearly formulated operational definitions that describe the behavioral issues of concern. Compared to structured interview methods, semi-structured interview strategies both depend on and foster greater interviewing skill. Like structured interview methods, semi-structured interview protocols require that all interview topics and questions are addressed at some point during an interview.

In the last stage of the pretest interview – following the free-narrative interview or semi-structured interview – the examiner will develop and review the test questions with the examinee (American Polygraph Association, 2009a, 2009b; Department of Defense, 2002). Test question language will be adjusted to ensure correct understanding and to account for information or admissions that the examinee may provide during the interview or while developing the test questions. Relevant questions will describe the possible behavioral involvement of the examinee in the issue or issues of concern. These questions will generally avoid issues related to memory, intent, and motivation. However, some investigative testing protocols will allow for test questions regarding memory if an examinee admits the alleged behavioral act and the issue of memory or motivation is target of the investigation (American Polygraph Association, 2009a).

When a polygraph examination consists of multiple series of test questions, the examiner will review each series of questions separately, then conduct the in-test data collection phase for each question series questions before reviewing and collecting data for each subsequent question series. When a polygraph consists of multiple series of test questions, there is no evaluation or discussion of the results of any individual series of questions until all test question series have been fully recorded and analyzed. If an acquaintance test was not conducted earlier it may be conducted after reviewing the test questions and before proceeding to the in-test phase of the exam. Some earlier polygraph testing formats employed a procedure analogous to an acquaintance test, though following the first presentation of the test stimuli during the in-test phase of the exam.

**In-test data collection**

The second phase of the polygraph examination is that of in-test data collection. This may be accomplished using any of a variety of validated diagnostic or screening test formats (American Polygraph Association, 2011b; Department of Defense, 2002). All screening and diagnostic polygraph techniques include relevant questions (RQs) that describe the examinee's possible involvement in the behavioral issues under investigation. Effective relevant questions will be simple, direct, and should avoid legal or clinical jargon and words for which the correct meaning may be ambiguous, confusing or not recognizable to persons unfamiliar with legal or professional vocabulary. Each relevant questions must address a single behavioral issue.

Relevant questions of event-specific diagnostic polygraphs are constructed with the assumption of non-independent criterion variance. The scientific and probabilistic meaning of this is that the RQs have a common or shared source of response variance because the external criterion states of different RQs may (and do) affect one another. The practical meaning of this is that all RQs must address behavior within a single incident of concern.

Multi-issue screening polygraphs, conducted in the absence of a known allegation or incident, may be constructed with relevant questions that describe distinct behaviors for which the external criterion states are assumed to vary independently (i.e., external criterion states are assumed to be exclusive or not interact and affect one another). There is evidence that response variance for these questions is not actually independent (Barland, Honts & Barger, 1989; Podlesny & Truslow, 1993; Raskin, Honts & Kircher, 2014), and for this reason field practices do not permit both positive and negative test results within a single examination. Regardless of whether conducted for diagnostic or screening purposes, all polygraph examinations are ultimately interpreted at the level of the test as a whole, though subtotal scores for individual RQs may be evaluated according to standardized procedures.

Most polygraph examinations in the United States today are conducted with some variant of the comparison question technique (CQT). The CQT was first described in publication by Summers (1939) while he was head of the Psychology Department at the Fordham University Graduate School in New York. The CQT was popularized within the polygraph profession by Reid (1947) and Backster (1963). It is the most commonly used and exhaustively researched family of polygraph techniques in use today. In addition to RQs, these polygraph techniques also include comparison questions (CQs; referred to in earlier polygraph literature as control questions). When scoring a test, examiners will numerically and statistically evaluate differences in responses to RQs and CQs.

The traditional form of comparison

question is the probable-lie comparison (PLC) questions, while some evidence-based contemporary CQTs make use of the directed-lie comparison (DLC). Examiners who use PLCs will maneuver the examinee into denying a common behavioral issue that is not the target of the examination. Probable-lie comparison questions have been the basis of some criticism of the polygraph technique due to their manipulative nature, and also the uncertainty surrounding the veracity of the examinee regarding these questions (Furedy, 1989; Lykken, 1981; Office of Technology Assessment, 1983; Saxe, 1991). Some of these criticisms rest on an inaccurate assumption that the polygraph measures actual lies *per se*. The polygraph, like many scientific tests, records responses to stimuli. Polygraph instruments do not actually measure *lies*, but instead discriminate deception and truth-telling through the use of probability models and statistical reference data that describe the differences in the patterns of reactions of truthful and deceptive persons when responding to RQs and CQs.

Although not a *control* in the strictest sense, CQs serve a similar function as a control in that they allow an examiner to effectively parse and compare diagnostic variance and other sources of variance. Response variance of CQs is not completely independent of the investigation target issue in the same way that scientific controls are - because responses for CQs and RQs are from the same examinee. This model of testing can be thought of as analogous to the way that data is acquired from the subjects in a two-way repeat measures ANOVA design - in which each subject serves as his or her own control set. In this way, each polygraph examination serves as a form of single subject scientific experiment.

Directed lie comparison (DLC) questions have been introduced as an alternative to the use of PLCs (Barland, 1981; Research Division Staff, 1995a; 1995b). DLCs are used in polygraph techniques developed by the United States (U.S.) government for use in polygraph screening programs, and in diagnostic polygraph techniques developed by researchers at the University of Utah (Honts & Raskin, 1988; Kircher, Honts & Raskin, 1997) and at the U.S. Department of Defense (Honts & Reavy, 2009). The major difference between PLC and DLC techniques is that DLC techniques are transparent and can be used without the need to maneuver or manipulate the examinee into denying a common behavioral issue.

DLCs have been shown in numerous studies, summarized by Blalock, Nelson, Handler and Shaw (2011; 2012), to perform classification tasks with equal efficiency and similar statistical distributions of numerical scores (American Polygraph Association, 2011) compared to PLC exams. Some researchers have suggested that DLCs are less ethically complicated than PLCs because they do not require the examiner to psychologically manipulate the examinee (Honts & Raskin, 1988; Honts & Reavy, 2009; Horowitz *et al.,* 1997; Kircher, Packard, Bell, & Bernhardt, 2001; Raskin & Kircher, 1990). DLC examinations have also been shown to retain effectiveness in different languages and cultures (Nelson, Handler & Morgan, 2012).

In addition to PLC and DLC questions, other variants of comparison questions have been suggested and argued, including exclusive comparison questions and non-exclusive (i.e., inclusive) comparison questions. Studies have shown all of these CQ variants to perform with similar effectiveness, for which accuracy does not differ at a statistically significant level (Amsel, 1999; Honts & Reavy, 2009; Horvath & Palmatier, 2008; Horvath, 1988; Palmatier, 1991). A recent meta-analytic survey (American Polygraph Association, 2011b) has further solidified this conclusion, in demonstrating that the same polygraph techniques perform with equivalent effectiveness and no significant differences in the sampling distributions of criterion deceptive and criterion truthful scores, when the techniques are employed with PLC or DLC questions. Scientific assumptions underlying scoring models for PLC and DLC techniques, assuming only that examinees will respond differently to relevant and comparison stimuli as a function of deception in response to RQs.

All polygraph techniques may include other procedural questions that are not numerically scored. Procedural questions designed for other technical examination purposes have not been supported by scientific studies, including: overall truth questions (Abrams, 1984; Hilliard, 1979), outside issue questions - referred to also as "symptomatic" questions - that attempt to inquire about interference from outside the scope of the examination questions (Honts, Amato & Gordon, 2004; Krapohl & Ryan, 2001), guilt-complex questions (Podlesny, Raskin & Barland, 1976), and sacrifice relevant questions regarding an examinee's intent to answer truthfully (Capps, 1991; Horvath, 1994). The absence of evidence to support their validity has led to the abandonment of the use of most of technical questions.

Only two non-scored technical questions remain widely used today, and these are used only in a procedural sense. Though not numerically scored, and not included in structured or statistical decision models, outside issue questions have been retained as a structural and procedural part of some test formats. Likewise, sacrifice questions are valued for the purported purpose of absorbing and discarding the examinee's initial response to the first question that describes the investigation target issue. These questions are also not numerically scored and not included in structured or statistical decision models. Un-scored sacrifice questions are included in virtually all modern polygraph techniques in use today.

A basic principle of measurement and testing is to obtain several measurements for each issue of concern. This is accomplished during polygraph testing by the use of several component sensors, each of which is designed to monitor increases or changes in activity in the autonomic nervous system, and by the standard practice of aggregating or combining the responses to several presentations of each test stimulus (Bell, Raskin, Honts, & Kircher, 1999; Kircher & Raskin, 1988; Raskin, Kircher, Honts, & Horowitz, 1988; Reid, 1947; Research Division Staff, 1995a; 1995b). Field polygraph procedures (Handler & Nelson, 2008; Kircher & Raskin, 1988; Department of Defense, 2006a) require that test stimuli are presented a minimum of three times and as many as five times. The common method is to repeat the entire series of test questions, while pausing the recording and deflating the cardio sensor in between repetitions. Some examination protocols (Department of Defense 1995a, 1995b; Handler, Nelson & Blalock, 2008) achieve several repetitions of the test questions without pausing the examination.

**Test data analysis – scoring of polygraph examinations**

Prior to informing the examinee or others of the results of the polygraph examination, the examiner must analyze the test data. Procedures for test data analysis are designed to partition and compare the sources of response variance: variance in response to RQs and variance in response to CQs. Responses are numerically coded and the result is compared to cutscores that represent normative expectations for deceptive or truthful persons. The overarching theory of polygraph testing is that responses to RQs and CQs vary significantly as a function of deception and truth-telling in response to the RQs.

The basic premise of numerical scoring of polygraph exams was first described by Kubis (1962) a method similar to that of Likert (1932) who showed how to reduce subjectivity using numerical coding of ordinal non-linear response data. Numerical scoring was popularized within the polygraph profession by Backster (1963) as the seven-position scoring system, and has been subject to further development and refinement through empirical study. The procedural construct for evaluation of differences in reaction to RQs and CQs can be traced to Summers (1939), who used a question sequence consisting of three relevant target questions and three comparative response questions repeated three times. Resulting data are found to produce different distributions for the different criterion groups, and these distributions can be used to classify other case observations. Variants of this model are observed in both signal detection and signal discrimination theory (Wickens, 1991; 2002).

There are three commonly used variants of the seven-position scoring system in use today, including the model developed by the U.S. Government (Department of Defense, 2006a; 2006b), the model published by ASTM International (2002), and the one developed by researchers from the University of Utah (Bell, Raskin, Honts, & Kircher, 1999; Kircher & Raskin, 1988; Raskin & Hare, 1978) and described by Handler (2006) and Handler and Nelson (2008). Differences between these seven-position methods are procedural and may be inconsequential in terms of test accuracy (American Polygraph Association, 2011b).

A commonly used modification of the seven-position scoring model is the three-position system defined by the Department of Defense (2006a, 2006b). Three-position scoring models are favored by some examiners for their simplicity and reliability, though there is a known increase in inconclusive test results (American Polygraph Association, 2011b) when using numerical cutscores intended for seven-position scores. The Empirical Scoring System (Nelson, Krapohl & Handler, 2008; Nelson & Handler, 2010; Nelson e*t al.,* 2011) is a statistically referenced, standardized, and evidence-based modification of the three-position and seven-position scoring models, with test accuracy comparable to other scoring models, though without the increase in inconclusive results.

As a theoretical matter, scoring of polygraph examinations is not different from the evaluation of other scientific tests in medicine, psychology, and forensics, and involves four basic concerns: 1) the identification of observable or measurable criteria (i.e., scoring features), 2) transformation of scoring features to numerical values, and reduction of numerical values to a grand total index for the examination as a whole and subtotal indices for the individual examination items, 3) statistical reference distributions to calculate statistical classifiers and numerical cutscores, and 4) structured decision policies.

Discussions of polygraph scoring methods are inseparable from discussions of test theory, including both decision theory and signal detection theory. Decision theory (Greenberg, 1982; Lehmann, 1950; Lieblich Ben-Shakhar, Kugelmass, & Cohen, 1978; Pratt, Raiffa, & Schlaifer, 1995; Wald, 1939;), like statistical learning theory (Hastie, Tibshirani & Friedman, 2001) is concerned with making optimal decisions. Signal detection theory is concerned with identifying and separating useful information from background noise or random information (Green, & Swets. 1966; Marcum, 1947; Schonhoff, & Giordano, 2006; Swets,1964; Swets, 1996; Tanner, Wilson, & Swets, 1954). Signal detection theory includes two fundamental models: signal detection (e.g., Yes or No) (Wickens, 2002) and signal discrimination models (e.g., A or B) (Wickens, 1991). CQT polygraph testing represents the second form, of these two, in that polygraph decisions attempt to achieve diagnostic accuracy when placing or predicting individual examinee membership into criterion categories of deception and truth-telling. Recent efforts (Nelson, Krapohl & Handler, 2008; Nelson *et al.,* 2011; Nelson & Handler, 2010) have begun to make more extensive use of statistical decision theory to quantify the probability of erroneous polygraph test results.

**Physiological reaction features.** Scoring of polygraph examinations begins with the identification of observable or measurable physiological responses that are correlated with deception and which can be combined into an efficient and effective diagnostic model. A number of studies - largely funded by the U.S. Department of Defense and conducted at the University of Utah and Johns Hopkins University - have described investigations into the identification and extraction of polygraph scoring features (Bell, Raskin, Honts, & Kircher, 1999; Harris, Horner, & McQuarrie, 2000; Kircher, Kristjiansson, Gardner, & Webb, 2005; Kircher & Raskin, 1988; Raskin *et al.*, 1988). These efforts are reflected in field practice standards published by the Department of Defense (2006a; 2006b), by ASTM International (2002), and in publications on the Empirical Scoring System (Nelson & Handler, 2010; Nelson *et al.,* 2011).

A small number of physiological indicators have repeatedly shown to be correlated with deception in structural decision models presently used in field polygraph programs. They are: respiration – observed as the respiration line length (Kircher & Raskin, 1988), respiration excursion length (Kircher & Raskin, 2002), or as either sustained decreases in respiration amplitude for three or more respiratory cycles, slowing of respiration rate for three or more cycles, or temporary increases in respiratory baseline of three cycles or more, and apnea; electrodermal activity – observed or measured as an increase in skin conductance (decrease in resistance), increased duration of response, and multiple responses; and cardiovascular activity – in the form of increase in relative blood pressure, increased duration of response, slowing of heart-rate, and decrease in finger blood-pulse volume.

Scoring features have been described as either primary or secondary (Bell, *et al.,*, 1999; Department of Defense, 2006a; 2006b). Primary features are those that capture the greatest degree of variance in deceptive and truthful responses to RQs and CQs within in each of the recorded physiological channels. Secondary features are correlated with differences in deceptive and truthful responses at statistically significant levels, but have weaker correlation coefficients compared to primary features. Also, secondary features provide information that is so strongly correlated with their primary counterparts that the added information is largely redundant and may not be additive to the effectiveness of some structural models. Some computerized scoring algorithms (Honts & Devitt, 1992; Kircher & Raskin, 1988; Krapohl, 2002; Krapohl & McManus, 1999; MacLaren & Krapohl, 2003; Nelson, Krapohl & Handler 2008; Raskin *et al.,* 1988;) in use today, and the evidence-based Empirical Scoring System (Nelson & Handler, 2010; Nelson *et al.,* 2011) have been designed to use only primary features, forgoing reaction features considered secondary in importance. Primary features, sometimes referred to as "Kircher features" are the following: respiration – observed as excursion length or correlated patterns, electrodermal activity - observed as the amplitude of vertical increase, and cardiovascular activity - observed as the amplitude of vertical increase in relative blood pressure.

**Numerical transformations.**

Numerical scores, in the form of non-parametric integers of positive or negative value, are assigned to each presentation of each RQs by comparing the strength of reaction to each RQ with the strength of reaction to the CQs presented in sequence with the RQs. A fundamental assumption during comparison question testing, is that both truthful and deceptive examinees may exhibit some degree of reaction to relevant questions stimuli. Indeed, Ansley (1999), Ansley and Krapohl (2000) and Offe and Offe (2007) have shown empirically that it is not the presence or absence of a response, nor the linear magnitude of response to the relevant questions that discriminates deception from truth-telling. Instead, the simple relative magnitude or degree of response to CQs, relative to the degree of response to the RQs, is the differentiating characteristic between deceptive and truthful examinees.

Deceptive examinees generally exhibit larger magnitude of change in autonomic activity in response to relevant stimuli than comparison stimuli, while truthful examinees will generally exhibit larger magnitude of change to comparison stimuli than to relevant stimuli. Deceptive scores are assigned when the magnitude of change to relevant question stimuli are greater than comparison question stimuli. Conversely, truthful scores are assigned whenever the degree of change in response to the comparison question stimuli is greater than responses to the relevant question stimuli. Numerous scientific reviews of countless scientific studies have affirmed the validity of the operational construct that responses to relevant and comparison stimuli vary as a function of deception or truth-telling regarding a past behavior (American Polygraph Association, 2011; Ansley, 1983; 1990; Abrams, 1973; 1977; 1989; National Research Council/National Academy of Science, 2003; Nelson & Handler, 2013; Office of Technology Assessment, 1983; Podlesny & Raskin, 1978; Raskin, Honts & Kircher, 2014).

**Decision cutscores and reference distributions.** Numerical test scores are translated into categorical test results through the comparison of test scores with cutscores that are anchored to reference distributions that describe the statistical density or probability of obtaining each particular score within the range of possible test scores. Because all scientific data are a combination of both diagnostic variance (i.e., explained variance) and unexplained variance (i.e., error variance, random variance or uncontrolled variance), individual scores can be expected to vary somewhat within each examination. For this reason, aggregated test scores have been found to provide the greatest diagnostic efficiency. This is often in the form of a grand total score, though subtotal scores are also used with some polygraph techniques. Grand total and subtotal scores are compared to cutscores and statistical reference distributions to determine the likelihood that an observed test scores has occurred due simply to uncontrolled error variance or random chance.

Probability cutscores are an expression of our tolerance for uncertainty or error, expressed as a statistical probability of error, often using the Greek letter α, declared prior to conducting an examination. A common probability cutscore in polygraph and other scientific disciplines is .05, with the goal of constraining the proportion of errors to 5% or less while attempting to provide a minimum confidence level of 95% for the categorical test result. Alternative probability boundaries of .10 and even .01, representing intended confidence levels of 90% and 99%, are sometimes used when testing objectives indicate a need for fewer inconclusive or unresolved test results or (.10) for fewer errors (.01).

Cutscores have also been determined using performance curves (Bell *et al.*, 1999) and through heuristic experience. Regardless of the method used to determine numerical cutscores, all decision cutscores will have some associated statistical information to describe the level of significance or probability of error. The relationship between numerical scores and associated statistical reference distributions can be calculated mathematically, and can also be conveniently determined using published reference tables.

Reference distributions have been summarized (American Polygraph Association, 2011; Nelson & Handler 2015) in the form of descriptive statistics that inform us about the location (i.e., mean or average), dispersion (i.e., variance or standard deviation) and shape of the distribution of scores observed in the sampling data for criterion deceptive and criterion truthful persons. Published reference distributions can be used to calculate the margin of uncertainty, in the form of a level of statistical significance, odds ratio, confidence level or probability of error, associated with any possible test score. Test results are said to be *statistically significant* when the probability of error is less than or equal to a declared probability cutscore or alpha level (i.e., $p <= \alpha$). This is equivalent to the condition when a test score equals or exceed a cutscore.

**Decision rules.** Decision rules are the practical and procedural comparison of test scores with either traditional cutscores (Bell, *et al.,* 1999; Department of Department of Defense, 2006a; 2006b; Kircher & Raskin, 1988; Raskin *et al.,* 1988) or cutscores selected for their level of statistical significance and probability of error using published reference distributions (American Polygraph Association, 2011, Krapohl & McManus, 1999; Krapohl, 2002; Nelson, Krapohl & Handler, 2008; Nelson & Handler, 2010; 2015; Nelson *et al.*, 2011). Procedurally, following the assignment of numerical scores, all scores are aggregated by summing the subtotal scores for all presentations of each RQ stimuli. Subtotal scores are then summed to achieve a grand-total score for event-specific diagnostic exams. Procedural decision rules are constructed with regard for assumptions of independent and non-independent criterion variance of the RQs of event-specific diagnostic exams and multi-issue screening exams.

Procedural decision rules for event specific diagnostic examinations, for which the criterion variance of the several RQs is assumed to be non-independent or dependent (i.e., all RQs address a single event for which the criterion status of different test stimuli may be strongly related), will make use of the grand total score to make the most accurate classification possible regarding the test as a whole. Some decision rules, such as those used by the Department of Defense (2006a; 2006b), as described by Light (1999), or those developed by Senter and Dollins (2002) may also make use of subtotal scores in attempt to reduce inconclusive results, increase test sensitivity to deception, or reduce false-negative errors. Decision results involving the grand total, referred to herein as the *grand-total-rule* (GTR), are accomplished by comparing the grand total score to cutscores for deceptive and truthful classifications. When using the GTR, test results are statistically significant and a categorical conclusion is made if the grand total score equals or exceeds one of the cutscores.

A two-stage modification of the GTR, referred to herein as the *two-stage-rule* (TSR) was described by Senter (2003) and Senter and Dollins (2002; 2008). The TSR allows the use of subtotal scores to achieve a categorical conclusion when the grand-total score is not statistically significant (i.e., inconclusive). When used, subtotal scores should be compared with cutscores that are statistically corrected for the known inflation of alpha, and associated potential increase in false-positive errors, that results from the use of multiple statistical comparisons (Abdi, 2007, Nelson and Handler, 2010; Nelson *et al.*, 2011; Nelson, Krapohl & Handler, 2008). Use of a simple statistical correction, referred to as a *Bonferroni correction*, can prevent an increase in false-positive errors when using subtotal scores of event-specific diagnostic exams.

By definition, the criterion states of the RQs of multi-issue screening exams are assumed to vary independently. For this reason, grand total scores are generally not used with multi-issue screening exams, for which the subtotal scores are more commonly used. Scores for multi-issue screening polygraphs are commonly evaluated using the individual subtotal scores (Department of Defense, 2006a; 2006b; Nelson and Handler, 2010; Nelson *et al.*, 2011; Nelson, Krapohl & Handler, 2008), using a rule referred to herein as the *subtotal-score-rule* (SSR). The SSR is executed by comparing each subtotal score to cutscores derived from statistical distributions of subtotal scores of examinations constructed of questions for which the criterion state was assumed to vary independently (American Polygraph Association, 2011; Nelson and Handler, 2010; Nelson *et al.*, 2011).

Although the SSR involves the use of individual subtotal scores, previous research (Barland, Honts & Barger, 1989; Podlesney & Truselow, 1993; Raskin, Honts & Kircher, 2014; Raskin, Kircher, Honts & Horowitz, 1988) has shown that although comparison question polygraph tests are effective at differentiating individuals who are truthful or deceptive, these tests are not as effective at determining the exact question or questions, within a series, to which an individual has lied or told the truth. The reasons for this may have to do with both the psychological and attentional demands of multiple independent stimulus targets, and the mathematical and statistical complexities that result from aggregating the sensitivity, specificity, false-positive and false-negative rates of multiple independent results. Test questions of multiple issue exams also have a shared, non-independent, source of response variance in the form of the examinee. For these reasons, the final classification of examination results as belonging to the groups of deceptive or truthful persons is always determined at the level of the test as a whole.

When using the SSR, the test result is classified as deceptive if *any* independent question produces a result that is statistically significant for deception, while truthful classifications require that the results of *all* independent questions are statistically significant for truth-telling. Field practices (American Polygraph Association, 2009a; Department of Defense, 2006a; 2006b) do not support the interpretation of responses to some questions as truthful and other responses as deceptive within a single examination. Of course, statistical methods involving regression, variance and covariance may provide capabilities not available within the simple procedural rubric of the SSR.

Subtotal cutscores for truth-telling should ideally be determined using procedures to statistically correct for the potential reduction of test specificity when requiring multiple statistically significant truthful scores before a truthful classification is made. A common solution for this correction in the statistical and mathematical sciences, as has been described in procedural methods for polygraph scoring (Nelson and Handler, 2010; Nelson *et al.*, 2011), is the *Šidák correction,* used when requiring multiple statistically significant independent probability events (Abdi, 2007). Some procedures involve the use of subtotal scores with traditional cutscores that are not derived from statistical reference distributions but are instead based on classification performance curves or heuristic experience (Bell *et al.,* 1999; Department of Defense, 2006a; 2006b).

**Physiological basis for the polygraph**

Although a thorough and detailed description of the physiological responses recorded by the polygraph is beyond the scope of this paper, a practical description of polygraph physiology in inextricably linked with the need to translate changes in recorded data into the form of test scores and test results. In contrast to earlier models for test data analysis that relied somewhat heavily on pattern recognition as a means of monitoring and observing physiological activity (Department of Defense, 2004), evidence-based models in use today will employ only those physiological features that are amenable to measurement and that have been shown, through published and replicated peer reviewed scientific studies, to be correlated at statistically significant levels with differences in response to different types of test stimuli that occur as a function of deception or truth-telling regarding past behavior (Bell, *et al.,* 1999; Harris, Horner & McQuarrie, 2000; Kircher, Krisjiansson, Gardner & Webb, 2005; Kircher & Raskin, 1988; Podlesny & Truslow, 1993, Raskin, Kircher, Honts & Horiwitz, 1988). Polygraph recording instrumentation has tended to focus on the acquisition of physiological response data that is of practical use to the task of scoring and interpreting polygraph test results, with few capabilities beyond that objective. For example: polygraph instrumentation is not used to evaluate cardiovascular or respiratory health.

Polygraph instrumentation consists minimally of three component sensors: two pneumograph sensors (thoracic and abdominal) to record breathing movement activity, electrical sensors to record autonomic activity in the palmar or distal regions (Handler, Nelson, Krapohl & Honts, 2010), and cardiovascular sensors to record relative changes in blood pressure (American Polygraph Association, 2009a, 2009b; Department of Defense 2006a). Vasomotor sensors (Kircher & Raskin, 1988; Bell *et al.,* 1999) are regarded as optional components of the polygraph instrument. Field testing protocols since 2007 have recommended the use of activity sensors to aid in the detection of countermeasure activity sensors, and are now required by the American Polygraph Association (2011a) as of January 1, 2012. This core combination of required sensors has been studied for several decades, and has been empirically shown to produce numerical scores that are structurally correlated with the criterion states of deception and truth-telling in statistical reference distributions from development and validation samples used in both field and laboratory studies (American Polygraph Association, 2011; Bell, Raskin, Honts & Kircher, 1999; Harris & Olsen, 1994; Harris, Horner & McQuarrie, 2000; Horowitz, Kircher, Honts & Raskin, 1997; Kircher & Raskin, 1988; Kircher, Kristjiansson, Gardner & Webb, 2005; MacLaren & Krapohl, 2003; Offe & Offe, 2007; Olsen, Harris & Chiu, 1994; Raskin, Kircher, Honts & Horowitz, 1988).

The physiological mechanics of polygraph responses during comparison question tests occur in the context of the autonomic nervous system (ANS) which includes both sympathetic (S/ANS) and parasympathetic (PS/ANS) components (Bear, Barry, & Paradiso, 2007; Costanzo, 2007; Maton et al., 1993; Paradiso, Bear, & Connors, 2007; Silverthorn, 2009; Standring, 2005). The ANS regulates involuntary processes including cardiac rhythm, respiration, salivation, perspiration, and other forms of arousal. S/ANS activity is responsible for stimulation of the internal organs in response to activity demands.

PS/ANS activity serves to reduce physiological activation to the minimum level necessary to ensure both longevity and adequate response to situational demands. PS/ANS and S/ANS activity are therefore in homeostatic balance with respect to real or perceived demands. The alternative to homeostatic balance is a general state of disease that may eventually lead to death. For this reason, every form of change in the ANS can be thought of as intended to maintain homeostasis and survival.

Polygraph examiners are primarily interested in recording and observing S/ANS activity, but it is important to understand that some activity, as with some cardiovascular data and respiration responses of interest to polygraph examiners, may actually be the result of changes in PS/ANS activity. (For more information about the dual innervation of the autonomic nervous system the reader is directed to more complete works by Janig (2006), Porges (2014), Handler and Richerter (2014)). The process of change, with the goal of maintaining homeostasis, is referred to as *allostasis* (Sterling & Eyer, 1988; Berntson & Cacioppo, 2007). Changes recorded during polygraph testing can be thought of as allostatic changes (Handler, Rovner & Nelson, 2008) that occur in an attempt to attain or maintain homeostasis.

Observable and recordable physiological changes in physiological activity that are structurally correlated with deception and truth-telling during comparison question testing include the following three features: 1) subtle and temporary respiratory suppression (i.e., suppression or reduction of respiratory movement, 2) relative magnitude of phasic electrodermal activity indicative of increased S/ANS activity, 3) relative magnitude of phasic response in the moving average of relative blood pressure. These measurable reactions have been described in several publications (ASTM International, 2002; Bell *et al.,* 1999; Department of Defense, 2006a, 2006b; Harris, Horner & McQuarrie, 2000; Kircher & Raskin, 1988; Kircher *et al.*, 2005; Krapohl & McManus, 1999; Raskin & Hare, 1978; Raskin *et al.,* 1988). Physiological responses for these three primary sensors (respiratory suppression, electrodermal activity, and cardiovascular activity) are easily observed and recorded.

Common misconceptions about the polygraph include the notion that the polygraph measures deep or rapid breathing, sweaty palms or sweating activity, and rapid or increasing heart rate activity. Of these, the first two, increased respiratory activity and sweating activity, are known to be inaccurate and unsatisfactory explanatory models for polygraph reactions. Only the third, heart rate, has been included in validated statistical classifiers for deception and truth-telling. However, it is *slowing* of cardiac rhythm, not increase, which is correlated with deception (Kircher *et al.,* 2005; Raskin & Hare, 1978). Pulse rate is included in the statistical model for the PolyScore algorithm (Blackwell, 1998; Dollins, Krapohl & Dutton, 1999; Dollins, Krapohl & Dutton, 2000; Harris & Olsen, 1994; Olsen *et al.*, 1991; Olsen *et al.*, 1994; Olsen, Harris, Capps & Ansley, 1997). Changes in heart rate are not included in other validated statistical models for scoring comparison questions tests. Pulse rate activity is therefore rarely included in polygraph decisions in field settings.

Manual analysis of the relative or absolute magnitude of change or response in phasic activity is easily accomplished for electrodermal activity and cardiovascular activity. Electrodermal data has been shown to be a strong indicator of S/ANS arousal (Boucsein, 2012), and to be the most robust and reliable contributor to the final score and resulting classification of comparison question polygraph test results (Ansley & Krapohl, 2000; Harris & Olsen, 1994; Kircher, 1981; 1983; Kircher & Raskin, 2002; Kircher & Raskin, 1988; Kircher *et al.*, 2005; Krapohl & McManus, 1999; Nelson, Krapohl & Handler, 2008; Olsen *et al.*, 1997; Raskin *et al.*, 1988).

Reactions to test stimuli can be evaluated through either non-parametric observation or through linear measurement. However, polygraph instrument manufacturer s have not completely standardized the signal processing and feature extraction methods whereby data obtained during polygraph testing are to be anchored to linear changes in physiological activity. As a result, measured or responses to polygraph stimuli are used only in automated statistical classification models within the polygraph testing paradigm and may not be directly related to measurements of similar physiological activity as utilized in medical fields. Most polygraph scoring paradigms will a non-parametric feature extraction method.

Cardiovascular responses during comparison question testing have been shown to be correlated with the criterion categories of deception and truth-telling at statistically significant levels (Bell *et al.*, 1999; Harris *et al.*, 2000; Kircher & Raskin, 1988; Kircher *et al.*, 2005; Nelson *et al.*, 2008; Raskin *et al.*, 1988). The structural correlation for cardiovascular response activity has been shown to be weaker than that of electrodermal response data, though stronger than that of respiratory response data. Diagnostic features and the interpretation of cardiovascular response activity were described by Handler and Reicherter (2008) and Handler, Geddes and Reicherter (2007). Some field polygraph examiners make use of photoelectric plethysmograph data, a form of cardiovascular recording for which information has been described by Handler and Krapohl (2007), Geddes (1974) and Honts, Handler, Shaw & Gougler (2015).

Of the three physiological sensors, respiratory data has found to be the most susceptible to disruption from voluntary activity during polygraph testing. Respiration data has the weakest structural coefficients of the required polygraph sensors (Harris & Olsen, 1994; Harris *et al.*, 2000; Kircher & Raskin, 1988; Kircher *et al.*, 2005; Nelson, Krapohl & Handler, 2008; Olsen *et al.*, 1997; Raskin *et al.*, 1988). However, field examiners have learned to evaluate respiration data for indicators of cooperation or non-cooperation during testing, in addition to evaluating respiration data for indicators of deception and truth-telling. Some research (Kircher *et al.*, 2005) has suggested that pneumograph data may be less diagnostic during comparison question tests conducted using DLC exams, while other findings have shown that respiration data of DLC exams does contain useable diagnostic information (Honts & Handler, 2014).

Respiratory suppression, though accurately measured by the curvilinear distance (Kircher & Raskin, 1988; Raskin *et al.*, 1988; Timm, 1982) or sum of absolute magnitude of change in y-axis excursion (Kircher & Raskin, 2002), is not easily measured without mechanical devices. Field polygraph examiners are taught to evaluate recorded data for the presence or absence of reaction patterns that have been described as correlated with the criterion categories of deception and truth-telling (Raskin & Hare, 1978; Bell *et al.*, 1999; Harris *et al.*, 2000; Kircher & Raskin, 1988; Kircher *et al.*, 2005; Raskin *et al.*, 1988). Pattern features shown to be correlated with respiratory suppression and CQT criterion categories are few, and include the following: 1) a subtle and temporary reduction of the tidal or inhalation volume resulting in a reduction of the y-axis (vertical) magnitude of the respiratory tracings for multiple respiratory cycles following the onset of the test question stimulus, 2) a subtle and temporary slowing of respiratory rate for multiple respiratory cycles following the onset of the test question stimulus, and 3) a subtle and temporary elevation of the exhalation baseline or residual volume for multiple respiratory cycles following the stimulus onset. Apnea is also correlated with differences in deception and truth-telling (Bell *et al.*, 1999; Kircher & Raskin, 1988), but can be easily feigned.

Polygraph sensors, while capable of recording sympathetic autonomic responses to test stimuli, are non-robust against disruptive somatic or physical activity that is sometimes not easily observed. In response to concerns about the potential for attempted faking during testing (i.e., countermeasures), somatic activity sensors have been developed to detect and record both overt and covert physical activity. There is indication in the literature that somatic activity sensors can increase examiners' ability to observe and detect these attempts (Ogilvie & Dutton, 2008; Stephenson & Barry, 1986). In the absence of recorded data of artifacted, odd or uninterpretable quality that indicates overt or covert physical activity, field examiners will assume that responses recorded by the respiration, electrodermal and cardiovascular sensors have their origins in the ANS and are not altered or contaminated by covert somatic activity.

**Psychological basis for the polygraph**

A satisfactory psychological theory will parsimoniously and holistically account for the variety of known and observed phenomena associated with the polygraph test. Such a theory will explain electrodermal responses, cardiovascular responses, and respiratory responses, to both PLC and DLC question, and will contribute to our understanding of test accuracy with both psychopathic and non-psychopathic persons. Moreover, a sound understanding of the psychological basis of the polygraph test will enable us to better understand issues of test suitability and unsuitability (i.e., for whom the test may or may not work). A comprehensive theoretical understanding of the psychological basis for responses to different testing paradigms such as the CQT and other polygraph and lie-detection paradigms – such as the concealed-information-test (CIT), which uses similar recorded physiological signals as a basis for ipsative calculations of the statistical significance of differences in responses to different test stimuli. Finally, a satisfactory psychological theory for polygraph testing will achieve a coherent integration of scientific knowledge regarding the polygraph with extant knowledge in related fields of science including cognitive, social and behavioral psychology, psychophysiology, signal detection theory, decision theory, statistical learning theory and more.

While a comprehensive discussion of the psychological basis for polygraph testing is beyond the scope of this paper, a brief explanation will hold that the psychological basis for responses to polygraph test stimuli involves a constellation of simple psychological mechanisms including cognition, emotion, and behavioral conditioning (Handler & Nelson, 2007; Handler, Shaw & Gougler, 2010; Kahn, Nelson, & Handler, 2009; Senter, Weatherman, Krapohl, & Horvath, 2010). Recently, preliminary process theory, related to orienting theory (Barry, 1996) has been suggested as a potentially parsimonious explanation for observed differences in response to different test stimuli (Palmatier & Rovner, 2014), though more discussion is needed to fully understand the advantages and limitations of this theory as applied to the polygraph. Until a more detailed evidence is described, a general constructed would suggest that all responses to test stimuli result from some combination of mental activity, emotion, and behavioral conditioning. All of these may play a role in physiological reactions that are load differentially for different types of polygraph test stimuli (i.e., relevant and comparison questions) as a function of deception or truth-telling in response to relevant stimuli that describe a behavioral issue of concern. It will be important to refrain from attempting to define which single emotion, or define the exact focus of attention and cognition within the examinee until such time as evidence exists to verify a more detailed description.

Field examiners have tended historically to simplify the explanation of polygraph psychology to a minimum level that satisfies both themselves and their examinees. This was often done using a scientifically unsatisfactory explanation of "psychological set" (see Handler & Nelson, 2007) as related to the fight-or-flight response that has been attributed to Cannon (1929). Although now regarded as an inadequate model for both polygraph responses and stress responses in general (Bracha, *et al.,* 2004; Taylor *et al.*, 2000), the application of this now problematic hypothesis holds that examinees will focus their attention and physiological response to the question or issue that presents the greatest immediate threat to their survival and well-being. The most obvious evidence of the limitations of the "psychological set" hypothesis is that it cannot account for the effectiveness of DLCs, and does not adequately explain test effectiveness with psychopaths who have been shown to have low levels of fear conditioning (Birbaumer, *et al.*, 2005). Additionally, the "psychological set" requires the assumption that polygraph sensors can identify different types of emotions, though the literature does not support this notion (Kahn, Nelson, & Handler, 2009). Moreover, this explanation suffers from a fundamental vulnerability to suggestions that it is pseudoscientific because it cannot satisfy a fundamental scientific requirement for falsifiability (Popper, 1959).

Handler & Nelson (2007) described the troublesome origins of the term "psychological set" which does not appear in the scientific psychological literature in the form employed by polygraph examiners. Differential salience has been suggested as a more general and parsimonious psychological theory that is more consistent with the field of scientific psychology, including emotion, cognition and conditioned learning as a basis of response to polygraph stimuli (Senter, Weatherman, Krapohl & Horvath, 2010).

Polygraph responses might also be accounted for using the conceptual framework of behavioral conditioning, as first described by Pavlov (1927), and learning theory, including the concepts of sensitization and habituation (Domjan, 2010, Groves & Thompson, 1970). A conditioned learning model for responses to polygraph stimuli suggests that involvement in a serious transgression amounts to a form of single-trial behavioral conditioning with test questions functioning as a conditioned stimulus. Polygraph interviewing theory holds that a thorough and effective pretest interview will give the truthful examinee an opportunity to habituate to test questions, while causing a deceptive examinee to become sensitized to the test questions as a conditioned stimulus.

Cognitive-behavioral theory, which includes cognition, emotion and behavioral/experiential learning as a basis of physiological response, has been also suggested as an explanatory hypothesis for the variety of known polygraph phenomena (Kahn, Nelson & Handler, 2009), and this model is consistent with the salience hypothesis described by Handler and Nelson (2007) and Senter *et al.,* (2010). A generalization of the cognitive-behavioral model for polygraph reactions suggests that truth-telling presents simpler cognitive and emotional task demands than deception.

A cognitive-behavioral and differential salience model would hold that physiological responses to a repeated sequence of polygraph test stimuli will be loaded onto different types of test stimuli as a function of deception and truth-telling regarding the investigation target issues, and that the basis of observed responses can be thought of as originating in cognition, memory, emotion and conditioned experience relative to the test stimuli. Relative differences in response to different types of test stimuli can be compared with statistical reference distributions and evaluated for their level of statistical significance to quantify the margin of uncertainty regarding a categorical conclusion of deception or truth telling. This form of theoretical explanation is fundamentally testable and therefore fundamentally scientific (Popper, 1959).

**Accuracy of polygraph examinations**

Results from several decades of scientific study have consistently supported the validity of the hypothesis that the combination of instrumental recording and statistical modeling can discriminate deception and truth-telling at rates significantly greater than chance. Scientific reviews of peer reviewed polygraph studies have borne this out repeatedly. Abrams (1989) surveyed the published literature and reported an accuracy level of .89. Honts and Peterson (1997), Raskin (2002), and Raskin & Podlesny (1979) reported the accuracy of polygraph studies as exceeding .90. The systematic review completed by the Office of Technology Assessment (1983) suggested that laboratory studies had an average unweighted accuracy of .83, with slightly higher accuracy, .85 from field studies at the time. Crewson (2001) reported an accuracy rate of .88 for diagnostic polygraphs in a comparison with medical and psychological tests. The National Research Council (2003) concluded with reservation that the polygraph differentiated deception from truth-telling at rates that were significantly greater than chance though less than perfect, and reported a median ROC of .89 for field studies and .86 for laboratory studies.

Different types of studies offer different advantages and disadvantages. Field studies offer assumed ecological validity, but are accompanied by a lack of experimental control, and by inconsistent case confirmation and non-random case selection – making generalization of some field study results troublesome or impossible. Laboratory studies offer the potential for random sampling and sufficient experimental control to study questions of causality, but have unknown ecological validity. The general trend in psychological research has been a high level of correspondence between field and laboratory studies (Anderson, Lindsay & Bushman, 1999), and this fact underscores the need to avoid confusing ecological validity with external validity.

External validity and ecological validity are not synonymous. External validity – the ability to generalize results to field settings - is often achieved from scientific studies in laboratory settings with imperfect ecological validity. Previous studies by Office of Technology Assessment (1983), the National Research Council (2003), the American Polygraph Association (2011), and Pollina *et al.*, (2004) showed no statistically significant differences in the results of polygraph test accuracy in field and laboratory studies.

The most recent scientific review of comparison question polygraph techniques in present use (American Polygraph Association, 2011) reported a mean accuracy of .89 for event-specific diagnostic polygraphs, with some evidence-based methods having been shown to provide mean accuracy levels in excess of .90. Multi-issue polygraphs, of the types used in operational security, law enforcement pre-employment, and post-conviction screening programs, have been shown to have a mean accuracy rate of .85. More important than mean accuracy statistics are the 95% confidence ranges that surround those point estimates, described later in this report, especially the lower-limit of test accuracy.

Earlier published scientific reviews (Abrams, 1973; Ansley, 1983; Ansley, 1990) have reported higher rates of test accuracy, often in the upper 90s. Results from these earlier systematic reviews are now thought to be confounded by sampling methodologies that may have overemphasized examiner self-report, possibly without researcher access to the recorded physiological data or numerical scores, and may have overemphasized the use of confession information as the case confirmation and selection criteria. These factors can potentially introduce non-random and non-representative case selection criteria that can systematically exclude both false-negative and false-positive error cases for which a confession is in not likely to be obtained. Advocacy research involving proprietary polygraph techniques seems to have also resulted in the production of exaggerated accuracy estimates – often near perfect – sometimes involving the principal investigator as examiner, scorer and technique developer/proprietor. Individual studies reporting near-perfect accuracy have been described as seriously methodologically flawed (American Polygraph Association, 2011).

As with all forms of scientific testing, diagnostic tests conducted in response to a single issue of concern, for which there is evidence of a problem, will provide greater overall accuracy than multi-issue exams (American Polygraph Association, 2011; Crewson, 2001) that are intended to simultaneously test several issues for which the criterion variance is assumed to be independent (i.e., a person could lie to one or more investigation target questions while not lying to other investigation targets). Multiple-issue examinations involve more probabilistic and statistical decisions, and therefore a greater aggregated potential for error and uncertainty compared to single issue exams. Other causes for differences in accuracy among diagnostic and screening exams may involve the competing attentional demands of multiple test target stimuli, and the potential that screening exam formats may at times be systematically biased for test sensitivity - with the goal of slightly over-predicting problems that can be resolved upon further investigation. It is also possible that some screening studies were completed using sub-optimal decision rules and cutscores that were not derived through scientific analysis.

The aggregated sensitivity rate for deception during diagnostic polygraphs was reported as .84 (95% confidence range .73 to .93) and the aggregated specificity rates for truth telling during diagnostic polygraphs was reported as .77 (95% confidence range .65 to .85) (American Polygraph Association, 2011). Evidence indicates that deceptive persons have a statistically significantly greater than chance probability of failing a polygraph test while truthful persons have a statistically significantly greater than chance probability of passing a polygraph.

Aggregated error estimates for polygraph diagnostic tests were calculated in the most recent meta-analytic survey using 24 peer reviewed scientific studies involving 8,975 confirmed scores of field and lab studies were reported as .08 for false negative errors (i.e., deceptive persons who pass the polygraph) and .12 for false positive errors (i.e., truthful persons who fail the polygraph). Inconclusive rates were reported as .09 for deceptive persons and .13 for truthful persons. The 95% confidence rage for decision accuracy of event-specific diagnostic polygraphs was .83 to .95 (American Polygraph Association, 2011). Quite obviously, deviations from empirically validated testing protocols may decrease expected test accuracy, and, of course, these accuracy estimates assume that each examination is conducted on a suitable examinee.

The aggregated sensitivity rate for deception during multi-issue polygraphs, of the type employed in polygraph screening programs, was reported as .77 (95% confidence range .60 to . 90) and the aggregated specificity rates for truth-telling during multi-issue criterion independent polygraphs was reported as .72 (95% confidence range .63 to .81) (American Polygraph Association, 2011). Evidence indicates that deceptive persons have a statistically significantly greater than chance probability of failing a test constructed of multiple independent issues, while truthful persons have a statistically significantly greater than chance probability of passing a similarly constructed multi-issue polygraph.

Aggregated error estimates for polygraph tests constructed from test questions for which the criterion variance is assumed to be independent, were calculated from 14 peer reviewed studies involving 1,194 confirmed scores of field and lab studies, were reported as .11 for false negative errors and .14 for false positive errors. Inconclusive rates were also reported as .11% for deceptive persons and .14 for truthful persons. The 95% confidence interval for unweighted average accurate rate for polygraph examinations constructed of test questions for which the criterion variance of the relevant questions was assumed to be independent was reported as .77 to .93 (American Polygraph Association, 2011).

For diagnostic polygraphs, the Positive Predictive Value (PPV), or probability that a failed polygraph result is correct, was reported as reported at .89 (95% confidence rage .81 to . 99, while the Negative Predictive Value (NPV), or probability that passed polygraph result is correct, was reported as .91 (95% confidence rage .82 to .99) (American Polygraph Association, 2011). For screening polygraphs, PPV was reported as .83 (95% confidence rage .71 to .94), while NPV was reported as .88, (95% confidence rage .78 to .97).

Conservative judgment necessitates the selection of the lower end of the confidence limit as the boundary at which we can be confident that polygraph accuracy exceeds. Therefore, diagnostic polygraphs can be assumed to provide accuracy over .81 for deceptive results, and over .82 for truthful results, while screening polygraphs can be assumed to provide accuracy over .71 for deceptive results and over .88 for truthful results. However, PPV and NPV are non-resistant to differences in base-rates, and figures reported herein apply only to balanced groups of polygraph exams. Common inferential estimates of test accuracy (e.g., sensitivity, specificity, inconclusive and error rates) are resistant to differences in base-rates and can be more useful when interpreting the meaning of the result of a single examination, such as when a court is evaluating an individual case.

**Threats to polygraph accuracy**

Because polygraph tests - like all tests - are inherently probabilistic (i.e., they are neither deterministic observation nor physical measurement), they are not perfect. No probabilistic test is completely immune to potential error or threats to test accuracy. Although the National Research Council (2003) could not find any scientific evidence that any personality type or endogenous factors significantly affect polygraph test accuracy, it is commonly understood that polygraph test accuracy may be compromised or reduced by the health, level of functioning or suitability of the examinee.

Abrams (1975) showed that polygraph test accuracy was reduced significantly with the level of functional maturity for young juveniles. In other studies, Abrams and Weinstein (1974) showed the polygraph cannot be expected to be accurate with subjects who have chronic mental health diagnoses within the psychotic spectrum of disorders, and further showed that polygraph accuracy is unstable for people whose intellectual abilities are below the lower limit of the normal range (Abrams, 1974).

Although developmental problems, low intellectual functioning, low functional maturity, and psychosis can adversely affect polygraph accuracy; there is no evidence that psychopathic personality issues will adversely affect polygraph test accuracy. Barland and Raskin (1975) studied criminal suspect with high psychopathic deviate scores on MMPI testing and showed no significant differences in the ability to detect deception. Patrick and Iacono (1989) also showed no significant differences in the detection of deception among psychopathic and non-psychopathic inmates. Raskin and Hare (1978) reported the same conclusion with a different sample of inmate subjects. Balloun and Holmes (1979) showed that polygraph accuracy using a guilty knowledge test paradigm was also not significantly different for college students with high and low psychopathic deviant scores on MMPI testing.

Although both the Office of Technology Assessment (1983) and the National Research Council (2003) expressed concern at the notion that polygraph test accuracy may be lower for persons with dangerous personality profiles, both reported that the published scientific evidence does not support, and consistently refutes, the hypothesis that psychopaths believe their lies and can therefore defeat the polygraph. In summary, polygraph testing with psychopathic persons can be assumed to be similar - as accurate and as inaccurate - as that with non-psychopathic persons. Regardless, public and media reactions may tend to simplistically assume that a person has “beaten” the polygraph whenever a testing error is observed. Some proportion of testing errors should not be surprising unless the proportion of errors can be shown as exceeding the 95% confidence interval for normally expected error rates.

In response to concerns that examinations conducted under *friendly* circumstances, such as those conducted under attorney-client privilege, have less validity than those conducted by law enforcement examiners, Honts and Peterson (1997) described flawed logic and reliance on a false hypothesis as the basis of this concern, and summarized the findings reported by Honts (1997) who investigated the hypothesis through logic, case analysis, and meta-analysis. At the present time there is no basis of evidence to support and available evidence contradicts the notion that exams conducted under attorney client privilege offer reduced accuracy or validity. These findings underscore the value of structured quantitative methods and the importance of objectivity and reproducibility when analyzing test data.

A final concern regarding polygraph test accuracy involves the possibility that countermeasures (i.e., faking) might be effective at altering the test outcome. This hypothesis represents an important concern for U.S. government and operational security programs, as well as for law enforcement pre-employment screening, and post-conviction supervision screening tests with convicted offenders. Despite the importance of this concern, only a small number of studies have been published on the topic of faking and polygraph accuracy.

Rovner (1979; 1986) and Rovner, Raskin and Kircher (1979) showed that having access to information about the polygraph technique was insufficient to significantly alter test accuracy. A concerning finding in these studies was that truthful subjects who attempted to employ countermeasures, in attempt to increase their assurance of passing, actually increased their likelihood of being classified as deceptive. These findings lead the National Research Council (2003) to conclude that countermeasure use by truthful examinees was not advisable.

Timm (1991) reported post-hypnotic suggestion to be ineffective as a polygraph countermeasure, while Ben-Shakhar and Dolev (1996) along with Elaad and Ben-Shakhar (1991) suggested that mental efforts may have an effect on the electrodermal channel primarily. Studies by Iacono, Boisvenu and Fleming (1984) and Iacono, Cerri, Patrick and Fleming (1992) showed benzodiazepines and stimulant medications to be ineffective countermeasures, though an earlier study by Waid, Orne and Orne (1981) indicated that meprobamate may adversely affect polygraph results. An inherent limitation to our acquisition of additional knowledge in this area will be that obtaining ethics committee approval to fully explore the effects of drugs or psychiatric medication on polygraph test results may be difficult or unlikely.

Concerning results regarding polygraph countermeasures have been described by Honts (1987), Honts, Amato & Gordon, (2004); Honts and Hodes (1983), Honts, Hodes and Raskin (1985), Honts, Raskin and Kircher (1987), and by Honts, Raskin, Kircher and Hodes (1988) whose collective work began to suggest that human polygraph experts are not as effective as they claim at differentiating countermeasure use from other artifacts. Additionally, Honts *et al.*, (1988) and Honts and Reavy (2009) found that spontaneous countermeasure use was not uncommon among both deceptive and truthful examinees. Raskin and Kircher (1990) reported that training in physical countermeasures can reduce polygraph test accuracy, and Honts, Raskin and Kircher (1994) reported mental and physical countermeasures as equally effective. In a different study, Honts, Winbush and Devitt (1994) reported that mental countermeasures can be used to defeat guilty knowledge tests. In another study, Honts and Amato (2001) again reported that countermeasures attempts by truthful subjects again resulted in the production of more deceptive test scores. Somatic activity sensors, and testing procedures intended to elucidate both mental and physical countermeasure, become more widely used following these studies, and replication of these studies is needed using contemporary testing instrumentation and methodologies.

Activity sensors are designed to be sensitive to somatic/behavioral nervous system activity while remaining robust against recording the effects of ANS activity of interest to polygraph test data analytic models. The rationale for activity sensors involves the fact that polygraph component sensors, while intended to be sensitive to sympathetic autonomic nervous system activity, are non-robust against also recording the effects of somatic activity. Stephenson and Barry, (1986) along with Ogilvie and Dutton (2008) showed that the addition of an activity sensor can increase the detection of somatic activity, and this may reduce the occurrence of false accusations of countermeasure use. It is assumed that observable activity indicates that the recorded polygraph data is likely to be an adulterated composite of both autonomic and somatic activity. Correspondingly, the absence of somatic activity would indicate the recorded autonomic data is most likely unaltered and authentic.

Statistical methods have not yet been widely exploited in countermeasure detection. However, the OSS-3 algorithm (Nelson *et al.*, 2008) includes a procedural requirement to review data for interpretable data quality and mark any segments of data that are artifacted by movement or other problem activity. The OSS-3 algorithm will aggregate the number and location of indicated artifact events and then calculate the statistical probability that observed artifacts have occurred due to random causes. The algorithm will alert the examiner to the possibility that an examinee may have attempted to systematically or intentionally alter the recorded physiological data whenever the likelihood falls below an established alpha boundary for statistical significance. Statistical methods have not yet been exhaustively studies, and additional research is needed regarding the application of statistical and computational methods to detect countermeasure attempts.

Strategies have been described involving both covert physical/muscle activity and mental activities. An alternative, social, countermeasure strategy will be an attempt to convince the examiner to ignore recorded indicators of deception as the result of some alternative cause.

Other forms of potential countermeasures may involve the use of medications or drugs, sleep deprivation or physical exhaustion, and the use of mental efforts such as hypnosis, meditation or metal activity. In general, polygraph countermeasures can be expected to attempt to either dampen or exaggerate responses to the entire set of test questions, resulting in an increased likelihood of an inconclusive test result. Alternatively, examinees may attempt to strategically alter responses to relevant or comparison test stimuli. Because it appears unwise to exaggerate responses to relevant stimuli, and because the reliable suppression of responses to only some test questions will present non-trivial complexities to the examinee, polygraph countermeasure strategies will most likely involve attempts to either dampen or disrupt the test as a whole or to augment or increase responses to comparison questions stimuli.

The goal of countermeasures or faking attempts, in terms of test data analysis, is to substantially alter the diagnostic and error variance contained in the recorded data such that it is uninterpretable or such there is no correlation between the deceptive or truthful criterion state and reaction differences that occur in response to different types of test stimulus questions. This would produce a test result that is inconclusive because the data are either uninterpretable or not statistically significant.

A sophisticated countermeasure objective would propose to alter the recorded test data such that the mathematical, statistical and computational methods to partition and compare the sources of variance would result in a direct reversal of the valence of the correlation coefficients for the criterion states of deception and truth-telling. Successful countermeasure attempts of this type would also require that the adulteration of both diagnostic and error variance in the recorded physiological data is accomplished in a manner that will convince trained examiners of the authentic quality of the recorded data. The inherent difficulty of this challenge is made more difficult by the fact that polygraph testing, as with other forms of testing, can include mechanisms and analytical procedures designed to quantify the probability that a persons has attempted to engage in countermeasures.

In response to the complexity of the issues, assertions and conclusions surrounding the potential for countermeasures during polygraph testing, the National Research Council (2003) wrote the following:

> "Because the effective application of mental or physical countermeasures on the part of examinees would require skill in distinguishing between relevant and comparison questions, skill in regulating physiological response, and skill in concealing countermeasures from trained examiners, claims that it is easy to train examinees to "beat" both the polygraph and trained examiners require scientific supporting evidence to be credible. However, we are not aware of any such research." (p.147).

The literature review of the National Research Council (2003) was unable to produce evidence supporting the hypothesis that polygraph countermeasures and faking attempts are effective at assisting deceptive persons to defeat comparison question polygraphs as they are presently used in field settings. However, this should not be interpreted as support of the infallibility of the polygraph, nor an assertion that nobody had ever passed a polygraph in error.

Although all test paradigms may have some potential vulnerability to exploitation, data at this time suggest that systematic efforts to alter the polygraph results are not well supported by evidence. Neither are claims that the polygraph test is infallible. Polygraph test results remain a probabilistic estimate of the degree of uncertainty surrounding a categorical conclusion. What is known at this time is that virtually all guilty or deceptive persons who agree to undergo polygraph testing may attempt to engage in some form of activity in an attempt to achieve a negative test result.

Polygraph has been shown to be an effective, even if imperfect, tool for discriminating deception and truth-telling even in field studies involving persons who were suspected of actual, sometimes serious, crimes - some of whom can be assumed to have attempted to engage in some form of countermeasure attempts to pass the test while lying. However, even a remote potential for effective countermeasure use - and the relationship of this to polygraph accuracy in government operational security, law enforcement pre-employment, and post-conviction supervision of convicted offenders in community settings - will mean that the interplay between countermeasure attempts and the authenticity of recorded test data will remain an important area of concern. Regardless of the effectiveness or ineffectiveness of countermeasures and faking attempts, additional and continuous research is needed to more fully understand the vulnerability to countermeasures of contemporary polygraph testing procedures.

**Contribution of polygraph results to professional decisions**

In an attempt to quantify the value and contribution of polygraph results to professional judgment, Honts and Schweinle (2009) described the use of the Information Gain Index (IGI, Wells & Olson, 2002) to diagnostic and screening polygraphs. IGI is a measurement of the increase in decision accuracy that results from method, and offers the advantage of providing information across the spectrum of prior base-rates – something other statistical metrics have failed to do efficiently. This method compares the increase in decision accuracy provided by polygraph results, with unassisted professional judgment – for which Vrij (2008) showed that police officers achieve a decision accuracy of 56% when attempting to determine deception or truth telling. The importance of the IGI statistic is that it provides information about the increase in decision accuracy across the entire range of possible base-rates.

Honts and Schweinle (2009) showed that diagnostic polygraph significantly increase the decision accuracy for both deceptive and truthful examinees, peaking at 27 times the accuracy of unassisted decisions, under assumed high base-rate conditions, with statistically significant increases in decision accuracy, compared to unassisted expert judgment, from base-rates .01 to .97 for deceptive outcomes and .03 to .99 for truthful outcomes. Screening polygraphs, often conducted under assumed low base-rate conditions, showed a statistically significant increase in the accuracy of decisions regarding deception, but no significant increase in the accuracy of decisions regarding truth telling. The increase in decision accuracy for deceptive outcomes was statistically significant, compared to unassisted lie-detection, from base-rates .02 to .83, and was statistically significant for truthful outcomes from base-rates .11 to .99.

Handler, Honts and Nelson (2013) further evaluated the IGI statistic using polygraph screening tests of the type used in operational security, law-enforcement screening, and post-conviction sex offender supervision programs. Handler *et al.*, showed statistically significant increases in the accuracy of screening decisions, compared to unassisted lie detection, from base-rates .01 to .94 for deceptive outcomes and from base-rates .07 to .99 for truthful outcomes. This suggests that effective use of polygraph testing, though probabilistic and imperfect, has the potential to increase the effectiveness of professional decision making.

## Conclusions

Evidence exists to support the scientific validity of polygraph testing in both diagnostic and screening contexts, and there is sufficient evidence to warrant continued interest in both research and practice of the instrumental and statistical discrimination of deception and truth telling in both forensic and screening programs. Available evidence can describe and account for virtually all aspects of the polygraph test, including general test theory, testing procedures, decision theory, and signal detection theory, test accuracy, and vulnerability to countermeasures or faking, along with the psychological and physiological basis of responses to polygraph stimuli.

The polygraph test, like other scientific tests, is a probabilistic test that involves the recording of physiological responses to stimuli and uses statistical decision theory to quantify the margin of error or level of statistical significance – or alternatively the odds or confidence level - associated with the test result (Nelson, 2014a; 2014b; 2014c, 2014d; 2014e). The polygraph test achieves its objectives through the structural combination of physiological responses that have been shown to be reliable proxies that are correlated at statistically significant levels with differences in responses that are loaded onto different types of test stimuli (i.e., RQs and CQs) as a function of deception and truth-telling regarding a past behavior.

The need for a test that can discriminate deception and truth-telling arises from the fact that evidence may not exist, or may not yet have been uncovered to enable a deterministic conclusion or physical measurement. A probabilistic test of deception – with accuracy significantly greater than both chance and unassisted lie detection - is the scientific alternative to the near chance accuracy rates of unaided human inference. Scientific tests are needed whenever a perfect deterministic observation is not possible.

Tests are often needed to make informed conclusions about events in the past, which can no longer be observed directly or deterministically, and also to understand the potential for future events which have not yet occurred and which therefore cannot yet be directly or deterministically observed. Polygraph test results refer to both the likelihood that a past behavior has occurred, and to the future potential that information or evidence will be uncovered to confirm a conclusion.

Scientific tests are also needed when there is a desire to measure an amorphous phenomena that cannot be subjected to physical measurement. Tests are not needed when direct mechanical or linear measurement is possible; in which case we simply measure the item of interest. Measurement, as compared to testing, involves mechanical measurement error. Testing of any type may involve human sources of variance and other random or uncontrolled sources of error variance in addition to the diagnostic variance contained within and expressed by the testing data.

All test data, including polygraph test data, are a combination of diagnostic variance (also referred to as explained variance, controlled variance, or signal) and error variance (also referred to as random variance, unexplained variance, uncontrolled variance, or noise). Ideally, testing data will include a large portion of diagnostic variance and a small portion of error variance, but no test is perfect. Scientific tests are expected only to quantify and account for the margin of error surrounding a test result, and to account for the basis of assumptions related to the testing procedures.

Because scientific tests are used to evaluate amorphous phenomena that cannot be subjected to deterministic observation or physical measurement, all test results are probabilistic – including when they are simplified to categorical test results. Quantification of the margin of error or level of statistical significance associated with a test result will enable referring professionals and consumers of testing results to make better-informed conclusions about the meaning and usefulness of a test result.

Concerns about the ethics of polygraph testing, and especially polygraph screening programs, have sometimes pointed to the lack of perfection and the false positive error rate as the basis for argument against the use of the polygraph. Expectations for deterministic perfection – in which test results are not affected by uncontrolled variance, human behavior, or random error – are not realistic, and frustration or disappointment regarding a lack of deterministic perfection are not warranted in a scientific testing context.

It is important to remember that one of the operational goals of testing - in medicine, psychology, forensics, polygraph programs and other testing contexts - is the reduction of harm resulting from both false-positive and false-negative errors. Any practical testing method that achieves accuracy significantly greater than chance has the potential for reducing such harm if used effectively. It is also important to note that screening tests of some types may be intended to slightly over-predict the presence of problems – with the goal of correcting errors with subsequent diagnostic testing. While false-positive errors can be identified and corrected with additional testing and investigation, identification of false negative errors is sometimes not possible until a problem has escalated to a degree that can sometimes permanently affect individual lives and futures. It is equally important to remember that neither polygraph test results, nor any form of test result, should be used alone as the basis for decisions that affect the rights and liberties of individuals. There are no published policies or standards of practice for screening polygraphs that advise or require the use of polygraph test results alone as a sufficient basis for professional decision-making.

As always, more information and research is desirable in some areas, pertaining to both theoretical constructs and practical concerns. Theories are satisfactory only as long as they account for known and observed phenomena. The emergence of any evidence or phenomena that is not accounted for by our current theories should be taken as an indicator of the need and obligation to continue to revise our knowledge and assumptions in response to the new information. Failure to make revisions to working theories is an indicator of stasis, and is a characteristic of pseudoscientific endeavors. Scientists are continuously upgrading all working theories in response to an ever-increasing volume of known and observed phenomena. All theories in the realm of science are expected to evolve over time and to move towards an integrated framework with other theories from other fields of science. For this reason, it will be important for the polygraph profession to continue to make use of new information from related fields of science.

Practical areas for which more information is needed include the accuracy of both diagnostic and screening polygraphs, the contribution of the results of diagnostic polygraphs to investigation outcomes, and the contribution of screening polygraphs to case and program outcome measures such as rule-violations, corruption, dereliction and recidivism. As long as some persons are motivated to engage in deception, and as long as others attempt to use scientific technologies to detect deception, there will be some who are interested in developing countermeasure strategies to evade detection. For this reason, there will be continued and ongoing interest in developing knowledge regarding the potential vulnerabilities of the polygraph test, and what additional methods can be applied to counter those vulnerabilities.

Though it is colloquially referred to as a "lie detector" test as a term of convenience, science and scientific reason do not suppose that the polygraph actually measures lies *per se*. All test results are probability statements. The alternative to a probabilistic understand-ding of polygraph test results is to encourage false expectations and frustration that the lies themselves can somehow be subjected to deterministic observation or physical measu-rement. In reality, the principles of physiology and psychology are sufficiently complex and variable that a probabilistic model is necessary and unavoidable.

Because lies *per se* are amorphous temporal events, lie detection will likely remain a probabilistic and imperfect task. It will be important to remain aware that the goals of scientific testing are often to quantify and measure phenomena that cannot be subjected to deterministic observation or mechanical measurement. Polygraph test results are a measurement of the uncertainty surrounding a categorical conclusion based on differences in responses that are loaded onto different types of test stimuli as function of deception or truth-telling regarding a behavioral concern.

**References**

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics.* Sage.

Abrams, S. (1973). Polygraph validity and reliability: A review. *Journal of Forensic Sciences*, *18*, 313-326.

Abrams, S. (1974). The validity of the polygraph with schizophrenics. *Polygraph, 3,* 328-337.

Abrams, S. (1975). A response to Lykken on the polygraph. *American Psychologist*, *30*, 709-711.

Abrams, S. (1977). *A polygraph handbook for attorneys.* Lexington, MA: Lexington Books.

Abrams, S. (1989). *The complete polygraph handbook.* Lexington, MA: Lexington Books.

Abrams, S. (. (1984). The question of the intent question. *Polygraph*, *13*, 326-332.

Abrams, S. & Weinstein, E. (1974). The validity of the polygraph with retardates. *Journal of Police Science and Administration*, *2*, 39766.

American Polygraph Association (2009a). *Model Policy for Post-conviction Sex Offender Testing.* [Electronic version] Retrieved January 25, 2012, from http://www .polygraph.org.

American Polygraph Association (2009b). *Model Policy for Law Enforcement Public Services Pre-Employment Screening Examinations* [Electronic version]1/25/2012, from http://www .polygraph.org.

American Polygraph Association (2011a). *By-laws: American Polygraph Association, effective 1-1-2012.* [Electronic version] Retrieved January 6, 2012, from http://www .polygraph.org.

American Polygraph Association (2011b). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph*, *40(4)*, 196-305.

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

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions In Psychological Science*, *8*, 3-9.

Ansley, N. (1983). A compendium on polygraph validity. *Polygraph*, *12*, 53-61.

Ansley, N. (1990). Law notes: Civil and criminal cases. *Polygraph*, *19*, 72-102.

Ansley, N. (1999, July). *The frequency of appearance of evaluative criteria in polygraph charts.* Defense Personnel Research Center.

Ansley, N. & Krapohl, D.J. (2000). The frequency of appearance of evaluative criteria in field polygraph charts. *Polygraph*, *29*, 169-176.

ASTM (2002). *Standard Practices for Interpretation of Psychophsysiological Detection of Deception (Polygraph) Data* (E 2229-02). ASTM International.

Backster, C. (1963). *Standardized polygraph notepack and technique guide: Backster zone comparison technique.* Cleve Backster: New York.

Balloun, K. D. & Holmes, D.S. (1979). Effects of repeated examinations on the ability to detect guilt with a polygraphic examination: A laboratory experiment with a real crime. *Journal of Applied Psychology*, *64*, 316-322.

Barland, G. H. (1981). *A Validity and Reliability Study of Counterintelligence Screening Test.* Fort George G. Meade, Maryland: Security Support Battalion, 902d Military Intelligence Group. [Reprinted in *Polygraph 41* (1) 1-27].

Barland, G. H., Honts, C. R. & Barger, S.D. (1989). *Studies of the accuracy of security screening polygraph examinations.* DTIC AD Number A304654. Department of Defense Polygraph Institute.

Barland, G. H. & Raskin, D.C. (1975a). An evaluation of field techniques in detection of deception. *Psychophysiology, 12* (3), 321-330.

Barland, G. H. & Raskin, D.C. (1975a). Psychopathy and detection of deception in criminal suspects. *Psychophysiology*, *12*, 224.

Barry, R. J. (1996). Preliminary process theory: Towards an integrated account of the psychophysiology of cognitive processes. *Acta Neurobiologiae Experimentalis*, *56(1*), 469-484.

Bear, M. F., Barry, W. C. & Paradiso, M.A. (2007). *Neuroscience*: Exploring the brain: Third Edition. Philadelphia, PA: Lippincott Williams & Wilkins.

Bell, B. G., Raskin, D. C., Honts, C. R. & Kircher, J.C. (1999). The Utah numerical scoring system. *Polygraph*, *28* (1), 1-9.

Ben-Shakhar, G. & Dolev, K. (1996). Psychophysiological detection through the guilty knowledge technique: Effects of mental countermeasures. *Journal of Applied Psychology*, *81*, 273-281.

Berntson, G. G., & Cacioppo, J. T. (2007). Integrative physiology: Homeostasis, allostasis and the orchestration of systemic physiology, in Cacioppo, J. T., Tassinary, L. G., and Berntson, G. G. (Eds.) *Handbook of Psychophysiology, 3rd edition* (pp. 433-449). New York: Cambridge University Press.

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W. & Flor, H. (2005). Deficient fear conditioning in psychopathy: a functional magnetic resonance imaging study. *Archives of general psychiatry, 62,* 799-805.

Blackwell, J. N. (1998). *PolyScore 33 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations*. Available at the Defense Technical Information Center. DTIC AD Number A355504/PAA. [Reprinted in *Polygraph, 28* (2), 149-175].

Blalock, B., Nelson, R., Handler, M. & Shaw, P. (2011). A position paper on the use of directed lie comparison questions in diagnostic and screening polygraphs. *Police Polygraph Digest,* (2011), 2-5.

Blalock, B., Nelson, R., Handler, M. & Shaw, P. (2012). The empirical basis for the use of directed lie comparison questions in diagnostic and screening polygraphs. *APA Magazine, 45(1),* 36-39.

Bracha, H. S., Ralston, T. C., Matsukawa, J. M., Williams, A. E. & Bracha, A.S. (2004). "Does "Fight or Flight" need updating? *Psychosomatics, 45* (5), 448–449.

Bradley, M. T. & Janisse, M.P. (1981). Accuracy demonstrations, threat, and the detection of deception: Cardiovascular, electrodermal, and pupillary measures. *Psychophysiology*, *18*, 307-315.

Boucsein, W. (2012). *Electrodermal activity.* New York: Plenum Press.

Campion, M., Campion, J. & & Hudson, J. (1994). Structured Interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79,* 998-1002.

Cannon, W. B. (1929). *Bodily changes in pain, hunger, fear, and rage.* New York: Appleton-Century-Croft

Capps, M. H. (1991). Predictive value of the sacrifice relevant. *Polygraph*, *20*, 1-6.

Costanzo, L. (2007). *Physiology*. Hagerstwon, MD: Lippincott Williams & Wilkins.Department of Defense (2004a). *Federal psychophysiological detection of deception examiner handbook.* [Retrieved from http://www.antipolygraph.org on 2-23-2008].

Crewson, P. E. (2001). *A comparative analysis of polygraph with other screening and diagnostic tools*. Research Support Service. Report No. DoDPI01-R-0003. Reprinted in *Polygraph 32* (57-85).

Department of Defense (2004b). *Test data analysis: DoDPI numerical evaluation scoring system.* [Retreived from http://www.antipolygraph.org on 6-28-2007].

Department of Defense (2006a). *Federal psychophysiological detection of deception examiner handbook*. Reprinted in *Polygraph, 40* (1), 2-66.

Department of Defense (2006b). *Test data analysis: DoDPI numerical evaluation scoring system.* [Retrieved from http://www.antipolygraph.org on 3-31-2007].

Department of Defense Polygraph Institute (2002). *Department of Defense Polygraph Institute Law Enforcement Pre-employment Test.* [Retrieved from http://www.antipolygraph.org on 6-2-2011].

Dollins, A. B., Krapohl, D. J. & Dutton, D. (1999). *A comparison of computer programs designed to evaluate psychophysiological detection of deception examinations: Bakeoff 1*. Department of Defense Polygraph Institute. Reprinted in *Polygraph 29* (237-247).

Domjan, M. (2010). *Principles of learning and behavior, 6th edition.* Cengage/Wadsworth.

Drever, E. (1995). *Using Semi-Structured Interviews in Small-Scale Research. A Teacher's Guide.* Scottish Council for Research in Education, Edinburgh.

Elaad, E. & Ben-Shakhar, G. (1991). Effects of mental countermeasures on psychophysiological detection in the guilty knowledge test. *International journal of psychophysiology: official journal of the International Organization of Psychophysiology*, *11*, 99-108.

Furedy, J. J. (1989). The North American CQT polygraph and the legal profession: a case of Canadian credulity and a cause for cultural concern. *The Criminal Law Quarterly, 31,* 431-451.

General Accounting Office (1991). *Using structured interviewing techniques.* Program Evaluation and Methodology Division [Retrieved online http://www.gao.gov/policy/10_1_5.pdf on 10-20-2009].

Green, D. M. & Swets. J A (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greenberg, I. (1982). The role of deception in decision theory. *Journal of Conflict Resolution, 26,* 139-156.

Groves, P. M. & Thompson, R.F. (1970). Habituation: A dual-process theory. *Psychological Review, 77*, 419-450.

Handler M. (2006). The Utah PLC. Polygraph, 35, 139-148.Handler, M. & Nelson, R. (2007). Polygraph terms for the 21st Century. *Polygraph*, *36*, 157-164.

Handler, M., Geddes, L., Reicherter, J. (2007). A discussion of two diagnostic features of the cardiovascular channel. *Polygraph, 36* (2)*,* 70-83.

Handler, M., Honts, C., Krapohl, D., Nelson, R. & Griffen, S. (2009). Integration of pre-employment polygraph screening into the police selection process. *Journal of Police and Criminal Psychology, 24*, 69-86.

Handler, M. & Krapohl, D. (2007). The use and benefits of the photoelectric plethysmograph in polygraph testing. *Polygraph, 36,* 18-25.

Handler, M. & Nelson, R. (2007). Polygraph terms for the 21st Century. *Polygraph, 36,* 157-164.

Handler, M. & Nelson, R. (2008). Utah approach to comparison question polygraph testing. *European Polygraph, 2,* 83-110.

Handler, M. Honts, C. & Nelson, R. (2013). Information gain of the Directed Lie Screening Test. *Polygraph, 42,* 192-202.

Handler, M., Nelson, R. & and Blalock, B. (2008). A focused polygraph technique for PCSOT and law enforcement screening programs. *Polygraph*, *37* (2), 100-111.

Handler, M., Nelson, R., Krapohl, J. & Honts, C. (2010). An EDA Primer for Polygraph Examiners. *Polygraph, 39*, 68-108.

Handler, M., & Reicherter, J. (2008). Respiratory blood pressure fluctuations observed during polygraph examinations. *Polygraph, 37* (4), 256-262.

Handler, M., Rovner, L. & Nelson, R. (2008). The concept of allostasis in polygraph testing. *Polygraph, 38,* 228-233.

Handler, M., Shaw, P. & Gougler, M. (2010). Some thoughts about feelings: A study of the role of cognition and emotion in polygraph testing. *Polygraph, 39,* 139-154.

Harris, J. C. & Olsen, D.E. (1994). Polygraph Automated Scoring System. Patent Number: 5,327,899. U.S. Patent and Trademark Office.

Harris, J., Horner, A. & McQuarrie, D. (2000). *An evaluation of the criteria taught by the department of defense polygraph institute for interpreting polygraph examinations*. Johns Hopkins University, Applied Physics Laboratory. SSD-POR-POR-00-7272.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *Elements of statistical learning.* Springer.

Hilliard, D. L. (1979). A cross analysis between relevant questions and a generalized intent to answer truthfully question. *Polygraph*, *8*, 73-77.

Hodgson, P. (1987). *A practical guide to successful interviewing.* Maidenhead: McGraw-Hill.

Honts, C. R. (1987). Interpreting research on polygraph countermeasures. *Journal of Police Science and Administration*, *15*, 204-209.

Honts, C. R. (1997). *Is it time to reject the friendly polygraph examiner hypothesis (FPEH)?* Paper presented at the American Psychological Society at the 9th annual meeting, Washington, D. C., May 23-26, 1997.

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

Honts, C. R. & Amato, S. (2001). Psychophysiological credibility assessment. *Journal of Forensic Psychology Practice*, *1*, 87-99.

Honts, C., Amato, S. & Gordon, A. (2004). Effects of outside issues on the comparison question test. *Journal of General Psychology*, *131* (1), 53-74.

Honts, C. R. & Devitt, M.K. (1992). B*ootstrap decision making for polygraph examinations.* Department of Defense Polygraph Institute report No DoDPI92-R-0002.

Honts, C. & Handler, M. (2014). Scoring respiration when using directed lie comparison questions. *Polygraph*, *43 (3)* 71-78.

Honts, C. Handler, M. Shaw, P. & Gougler, M. (2015). The vasomotor response in the comparison question test. *Polygraph 44* (1) 62-78.

Honts, C. R. & Hodes, R.L. (1983). The detection of physical countermeasures. *Polygraph*, *12*, 7-17.

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

Honts, C. R. & Peterson, C.F. (1997). *Brief of the Committee of Concerned Social Scientists as Amicus Curiae United States v Scheffer*. Available from the author.

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

Honts, C. R., Raskin, D. C. & Kircher, J.C. (1987). Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Psychophysiology*, *1*, 241-247.

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

Honts, C. R., Raskin, D. C., Kircher, J. C. & Hodes, R.L. (1988). Effects of spontaneous countermeasures on the physiological detection of deception. *Journal of Police Science and Administration*, *16*, 91-94.

Honts, C. R. & Reavy, R. (2009). *Effects of comparison question type and between test stimulation on the validity of comparison question test. Final progress report on contract No.W911Nf-07-1-0670*, submitted to the Defense Academy of Credibility Assessment (DACA). Boise State University.

Honts, C. R., Winbush, M. & Devitt, M.K. (1994). Physical and mental countermeasures can be used to defeat guilty knowledge tests. *Psychophysiology*, *31*, S57.

Horneman, C. J. & O'Gorman, J.G. (1985). Detectability in the card test as a function of the subject's verbal response. *Psychophysiology*, *22*, 330-333.

Horowitz, S. W., Kircher, J. C. & Raskin, D.C. (1986). Does stimulation test accuracy predict accuracy of polygraph test?. *Psychophysiology*, *23*, 442.

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. & Palmatier, J. (2008). Effect of two types of control questions and two question formats on the outcomes of polygraph examinations. *Journal of Forensic Sciences*, *53(4)*, 1-11.

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

Horvath, F. S. (1994). The value and effectiveness of the sacrifice relevant question: An empirical assessment. *Polygraph*, *23*, 261-279.

Iacono, W. G., Boisvenu, G. A. & Fleming, J.A. (1984). Effects of diazepam and methylphenidate on the electrodermal detection of guilty knowledge. *Journal of Applied Psychology*, *69*, 289-299.

Iacono, W. G., Cerri, A. M., Patrick, C. J. & Fleming, J.A. (1992). Use of antianxiety drugs as countermeasures in the detection of guilty knowledge. *The Journal of applied psychology*, *77*, 60-4.

Janig. W. (2006). *The Integrative Action of the Autonomic Nervous System: Neurobiology of Homeostasis.* Cambridge University Press.

Kahn, R. L. and C. F. Cannell (1957). *The psychological basis of the interview. The dynamics of interviewing: theory, technique, and cases.* New York, John Wiley & Sons: 22-64.

Kahn, J., Nelson, R. & Handler, M. (2009). An exploration of emotion and cognition during polygraph testing. *Polygraph*, *38*, 184-197.

Kirby, S. L. (1981). The comparison of two stimulus tests and their effect on the polygraph technique. *Polygraph*, *10*, 63-76.

Kircher, J. C. (1981). *Computerized chart evaluation in the detection of deception.* Masters thesis: University of Utah.

Kircher, J. C. (1983). *Computerized decision making and patterns of activation in the detection of deception.* Dissertation Abstracts International, 44, 345.

Kircher, J. & Raskin, D. (2002). Computer methods for the psychophysiological detection of deception. In Murray Kleiner (Ed.), *Handbook of Polygraph Testing* . San Diego: Academic Press.

Kircher, J. C., Kristjiansson, S. D., Gardner, M. K. & Webb, A. (2005). *Human and computer decision-making in the psychophysiological detection of deception.* University of Utah.

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

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

Krapohl, D. J. (2002). Short report: Update for the objective scoring system. *Polygraph, 31,* 298-302.

Krapohl, D. & McManus, B. (1999). An objective method for manually scoring polygraph data. *Polygraph*, *28*, 209-222.

Krapohl, D. J. & Ryan, A.H. (2001). Final Comment on the bleated look at symptomatic questions. *Polygraph*, *30*, 218-219.

Kvale, S. (1996). *Interviews: An Introduction to Qualitative Research Interviewing.* Sage Publications.

Lehmann, E. L. (1950). Some principles of the theory of testing hypotheses. *Annals of Mathematical Statistics, 21* (1)*,* 1-26.

Lieblich, I., Ben-Shakhar, G., Kugelmass, S. & Cohen, Y. (1978). Decision theory approach to the problem of polygraph interrogation. *Journal of Applied Psychology, 63,* 489-498.

Light, G. D. (1999). Numerical evaluation of the Army zone comparison test. *Polygraph, 28*, 37-45.

Lindlof, T. R. & Taylor, B.C. (2002). *Qualitative communication research methods*. Thousand Oaks: Sage.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 5-55.

Lykken, D. T. (1981). *A tremor in the blood: Uses and abuses of the lie detector*. New York:McGraw-Hill.

MacLaren, V. & Krapohl, D. (2003). Objective Assessment of Comparison Question Polygraphy. *Polygraph*, *32*, 107-126.

Marcum, J. I. (1947). *A statistical theory of target detection by pulsed radar.* Rand Corporation. [Retrieved online http://www.rand.org 2011-11-2].

Maton, A., Hopkins, J., McLaughlin, C., Johnson, S., Quon Warner, M., LHart, D. & Wright, J. (1993). *Human biology and health.* Englewood Cliffs, New Jersey, USA: Prentice Hall.

National Research Council (2003). *The polygraph and lie detection.* Washington, D.C.: National Academy of Sciences.

Nelson, R. (2014a). What does the polygraph measure? *APA Magazine, 47(2)* 39-47.

Nelson, R. (2014b). Redux: What does the polygraph measure? (in 600 words or less). *APA Magazine, 47* (3)*,* 36-37.

Nelson, R. (2014c). Take 3: What does the polygraph measure? (in 250 words or less). *APA Magazine, 47* (4), 60.

Nelson, R. (2014d). Short Answer: What does the polygraph measure? (in 150 words or less). *APA Magazine, 47* (5), 27.

Nelson, R. (2014e). Sound-bite: What does the polygraph measure? (in 50 words or less). *APA Magazine, 47* (6), 29.

Nelson, R. & Handler, M. (2010). *Empirical scoring system: NPC quick reference.* Lafayette Instrument Company. Lafayette, IN.

Nelson, R. & Handler, M. (2012). Monte Carlo study of criterion validity of the directed lie screening test using the Empirical Scoring System and the Objective Scoring System version 3. *Polygraph*, *41,* 145-155.

Nelson, R. & Handler, M. (2013). A brief history of scientific reviews of polygraph accuracy research. *APA Magazine, 47* (6), 22-28.

Nelson, R, & Handler, M. (2015,). Statistical reference distributions for comparison question polygraphs. *Polygraph, 44* (1), 91-114.

Nelson, R., Handler, M., Blalock, B. & Hernandez, N. (2012). Replication and extension study of Directed Lie Screening Tests: criterion validity with the seven and three Position models and the Empirical Scoring System. *Polygraph 41* (3), 186-198.

Nelson, R., Handler, M., & Morgan, C. (2012). Criterion validity of the directed lie screening test and the empirical scoring system with inexperienced examiners and non-naive examinees in a laboratory setting. *Polygraph 41* (3), 176-185.

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B. & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph*, *40*, 67-78.

Nelson, R., Krapohl, D. & Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. *Polygraph*, *37*, 185-215.

Offe, H. & Offe, S. (2007). The comparison question test: does it work and if so how?. *Law and human behavior*, *31*, 291-303.

Office of Technology Assessment (1983). *The validity of polygraph testing: A research review and evaluation*. Printed in *Polygraph, 12,* 198-319.

Ogilvie, J. & Dutton, D. (2008). Improving the detection of physical countermeasures with chair sensors. *Polygraph*, *37* (2), 136-148.

Olsen, D. E., Ansley, N., Feldberg, I. E., Harris, J. C. & Cristion, J.A. (1991). Recent developments in polygraph technology. *Johns Hopkins APL Technical Digest*, *12*, 347-357.

Olsen, D. E., Harris, J. C. & Chiu, W.W. (1994). The development of a physiological detection of deception scoring algorithm. *Psychophysiology*, *31*, S11.

Olsen, D. E., Harris, J. C., Capps, M. H. & Ansley, N. (1997). Computerized polygraph scoring system. *Journal of Forensic Sciences*, *42*, 61-70.

Palmatier, J. J. (1991). *Analysis of two variations of control question polygraph testing utilizing exclusive and nonexclusive controls*. Unpublished doctoral dissertation.

Palmatier, J., Rovner, L (2014). Credibility assessment: Preliminary process theory, the polygraph process, and construct validity. *International Journal of Psychophysiology,* [available online: http://www.ncbi.nlm.nih.gov/pubmed/24933412].

Paradiso, M. A., Bear, M. F. & Connors, B.W. (2007). *Neuroscience: Exploring the Brain.* Hagerstwon, MD: Lippincott Williams & Wilkins.

Patrick, C. J. & Iacono, W.G. (1989). Psychopathy, threat and polygraph test accuracy. *Journal of Applied Psychology*, *74*, 347-355.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* Sage Publications Inc. 2009.

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

Podlesny, J., Raskin, D. & Barland, G. (1976). *Effectiveness of techniques and physiological measures in the detection of deception.* Report No. 76-5, Contract 75-N1-99-001 LEAA (available through Department of Psychology, University of Utah, Salt Lake City).

Popper, K. R. (1959). *The logic of scientific discovery.* Routledge.

Porges, S. (2011). *The Polyvagal Theory: Neurophysiological Foundations of Emotions, Attachment, Communication, and Self-regulation.* Norton.

Pratt, J., Raiffa, H. & Schlaifer, R. (1995). *Introduction to statistical decision theory.* Cumberland, RI: MIT Press.

Pollina, D. A., Dollins, A. B., Senter, S. M., Krapohl, D. J. & Ryan, A. H. (2004). Comparison of polygraph data obtained from individuals involved in mock crimes and actual criminal investigations. *Journal of applied psychology, 89,* 1099-105.

Popper, K. R., (1959). *The Logic of Scientific Discovery.* The German version is currently in print by Mohr Siebeck (ISBN 3-16-148410-X), the English one by Routledge publishers (ISBN 0-415-27844-9).

Powell, M. B. & Snow, P.C. (2007). Guide to questioning children during the free-narrative phase of an investigative interview. *Australian Psychologist*, *42* (1), 57-65.

Raffle, A. E. & Muir Gray, J.A. (2007). *Screening.* Oxford University Press.

Raskin D. C., Honts. C. R. (2002). The comparison question test. In M. Kleiner (Ed.), *Handbook of Polygraph Testing*. San Diego:Academic Press.

Raskin, D. C., Honts, C. R., Kircher, J. C. (2014). *Credibility assessment: Scientific research and applications.* Academic Press.

Raskin, D. C. & Hare, R.D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology*, *15*, 126-136.

Raskin, D. C. & Kircher, J.C. (1990, May 2). *Development of a computerized polygraph system and physiological measures for detection of deception and countermeasures: A pilot study.* Scientific Assessment Technologies, Inc.

Raskin, D. C. & Podlesny, J.A. (1979). Truth and deception: A reply to Lykken. *Psychological Bulletin*, *86*, 54-59.

Raskin, D., Kircher, J. C., Honts, C. R. & Horowitz, S.W. (1988). *A study of the validity of polygraph examinations in criminal investigations.* Final Report, National Institute of Justice, Grant No. 85-IJ-CX-0040.

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law and Criminology*, *37*, 542-547. Reprinted in *Polygraph 11*, 17-21.

Research Division Staff (1995a). *Psychophysiological detection of deception accuracy rates obtained using the test for espionage and sabotage.* DTIC AD Number A330774. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in *Polygraph, 27*, (3), 171-180.

Research Division Staff (1995b). *A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope Polygraph and the test for espionage and sabotage question formats.* DTIC AD Number A319333. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in *Polygraph, 26* (2), 79-106.

Richerter J. & Handler, M. (2014). *A physiology manual for PDD lifelong learners of the science.* Available from the authors. Retrieved online at [http://www.polygraph.org/files/apa_psychophysiology_study_guide.pdf] on 11-30-2014.

Rovner, L. I. (1979). *The effects of information and practice on the accuracy of physiological detection of deception.* Unpublished doctoral dissertation: University of Utah.

Rovner, L. I. (1986). Accuracy of physiological detection of deception for subjects with prior knowledge. *Polygraph*, *15* (1), 1-39.

Rovner, L. I., Raskin, D. C. & Kircher, J.C. (1979). Effects of information and practice on detection of deception. *Psychophysiology*, *16*, 197-198 (abstract).

Standring, S. (2005). *Gray's anatomy* (39th ed.). Elsevier Churchill Livingstone.

Saxe, L. (1991). Science and the CQT polygraph: A theoretical critique. *Integrative Physiological and Behavioral Science, 26*, 223-231.

Schonhoff, T. A. & Giordano, A.A. (2006). Detection and estimation theory and its applications. New Jersey: Pearson Education.

Senter, S. M. (2003). Modified general question test decision rule exploration. *Polygraph, 32*, 251-263.

Senter, S. M. & Dollins, A.B. (2002). *New decision rule development: Exploration of a two-stage approach. Report number DoDPI00-R-0001.* Department of Defense Polygraph Institute Research Division, Fort Jackson, SC.

Senter, S. M. & Dollins, A.B. (2008). Optimal decision rules for evaluating psychophysiological detection of deception data: an exploration. *Polygraph, 37* (2), 112-124.

Senter, S., Weatherman, D., Krapohl, D. & Horvath, F. (2010). Psychological set or differential salience: A proposal for reconciling theory and terminology in polygraph testing. *Polygraph, 39* (20), 109-117.

Silverthorn, D. U. (2009). *Human physiology: An integrated approach* (4 ed.). Pearson/Benjamin Cummings.

Stephenson, M. & Barry, G. (1986). Use of a motion chair in the detection of physical countermeasures. [Reprinted in *Polygraph, 17*, 21-27].

Sterling, P. and Eyer, J. (1988) Allostatsis: a new paradigm to explain arousal pathology. In: Fisher, S., Reason, J. (Eds.) *Handbook of Life Stress, Cognition and Health.* New York: Wiley and Sons.

Summers, W. G. (1939). Science can get the confession. *Fordham Law Review, 8,* 334-354.

Swets, J. A. (1964). Signal detection and recognition by human observers. New York: Wiley.

Swets, J. A. (1996). Signal detection theory and ROC analysis in psychology and diagnosis. Mahwah, NJ: Erlbaum.

Tanner, J., Wilson, P. & Swets, J. (1954). A decision-making theory of visual detection. *Psychological Review, 61* (6), 401-409.

Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A. & Updegraff, J.A. (2000). Biobehavioral responses to stress in females: Tend-and-befriend, not fight-or-flight. *Psychological Review, 107*, 411-429.

Timm, H. W. (1982). Analyzing deception from respiration patterns. *Journal of Police Science and Administration*, *10*, 47-51.

Timm, H. W. (1991). Effect of posthypnotic suggestions on the accuracy of preemployment polygraph testing. *Journal of Forensic Sciences*, *36*, 1521-1535.

Vrij, A. (2008). *Detecting Lies and Deceit, Pitfalls and Opportunities, 2nd edition*. West Sussex, Enland: Wiley & Sons.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics, 10* (4), 299-326.

Waid, W. M., Orne, E. C. & Orne, M.T. (1981). Selective memory for social information, alertness, and physiological arousal in the detection of deception. *Journal of Applied Psychology*, *66*, 224-232.

Wells, G. L. & & Olson, E.A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 8*, 155–167.

Wickens, T. D. (1991). Maximum-likelihood estimation of a multivariate Gaussian rating model with excluded data. *Journal of Mathematical Psychology*, *36*, 213-234.

Wickens, T. D. (2002). Elementary signal detection theory. New York: Oxford.

Widup, R. & Barland, G.H. (1994). *Effect of the location of the numbers test on examiner decision rates in criminal psychophysiological detection of deception tests*. Department of Defense Polygraph Institute.

Wilson, J. & Jungner, G. (1968). Principles and practice of screening for disease. *WHO Chronicle Geneva: World Health Organization*, *22* (11):473.