Polygraph Principles: A Literature Review

Donald J. Krapohl[1]

Abstract

Since the emergence of polygraphy as a field practice in the first half of the 20th century the discipline has been beset with internal debates over which methodologies were the best. This contest is better understood in the context of how polygraphy evolved. Polygraph practitioners became the pioneers who would chart the course for the profession. Virtually none had educational preparation for test development. All of these pioneers were in private practice, all had commercial polygraph schools, and all developed methodologies which they aggressively promoted and taught in their individual schools. Consequently, the debate about which methods were "best" was inextricably tied to entrenched economic interests. Polygraph research began in earnest in the 1970s, some decades after many of the schools had staked out their territories. Today, nearly all legacy techniques have components that have since been borne out in research; nearly all of them included erroneous elements that are the product of bias, self-interest, and often naïveté regarding psychometrics, psychophysiology, and decision theory. This paper sets out to identify and summarize 20 separate polygraph principles based on published research that transcend any particular polygraph technique. Awareness of these principles may be beneficial to professional examiners sorting through the claims of polygraph authorities, and help in the selection and execution of their polygraph practices.

Introduction

Let us begin with an audaciously condensed summary of the polygraph profession's first 60 years from a slightly different perspective: the history of the interaction of polygraph profession and the relevant scientific community, and where it ultimately led.

As every polygraph student knows, the birth of the polygraph "lie detector" is placed in the 1920s with the work of John Larson and Leonarde Keeler. Their successes in crime solving led to the creation of the polygraph field. The field grew geometrically from that point for decades, and the polygraph would eventually be found in almost every conceivable application. Despite polygraphy's rapid growth, scientists were slow to arrive on the scene to help sift through what could be supported from what should be abandoned or avoided. That information gap was relegated to polygraph "experts" who promoted their own ideas. It would not be until the 1970s that the polygraph attracted any more than intermittent scientific attention.

Parochial and economic interests prevailed in the field, even as research began to appear. The considerable value placed on these interests contributed to invalid, and in a few cases, unethical polygraph practices. Public opinion became aroused by these problems, and by the 1960s calls for change were being heard from many sides. Rather than adopting practices that were more defensible, polygraph examiners, in the form

[1] The views expressed in this article are those of the author, and do not necessarily represent those of the Department of Defense or the US Government. The author is grateful to Keith Gaines, Mark Handler and Donnie Dutton for their comments to an earlier draft of this paper. The article is one in a continuing series under the general title "Best Practices." Requests for reprints can be sent to APAkrapohl@gmail.com.

Additional disclaimer: Portions of the research herein may be at odds with policies of some department or agencies. This is sometimes unavoidable as research continues to improve our understanding of polygraphy. Polygraph examiners should advocate for alignment of policies with evidence, but avoid unilaterally departing from policy.

of associations, circled the wagons against the growing public clamor. Not only was nearly any testing method tolerated by the profession organizations, but some of the most irredeemable practices were condoned (One government examiner published an article to defend his agency's practice of seven-second question spacing). With public sentiment stirred by what polygraph examiners were doing, the US Congress intervened in 1988 with the Employee Polygraph Protection Act (EPPA) to provide safeguards to the public that the then-trade associations had failed to do.

Lessons Learned

It has been a quarter century since the EPPA came to be, and many things have changed. Most examiners are now coming from the public sector rather than private practice, and with this shift has come a new zeitgeist. One significant departure from the pre-EPPA collective mindset, at least for the American Polygraph Association, has been the decidedly deliberate steps to move away from the almost-singular focus on the advancement of the interests of the practitioner (e.g. trade or industry association) that helped bring about EPPA, to a more balanced and enlightened perspective of also ensuring their members delivered valid and responsible services to the public (professional association). It is a most promising sign. This adaptation was long in coming, and the observation that the transformation is taking place is based on four events:

- The Association now holds members accountable for the validity and reliability of their chosen methods.
- The Association has an educational initiative to help practitioners select and properly use valid and reliable methods.
- The Association has assembled a body of best practice guides for a variety of applications.
- The Association's public relations effort includes conveying those best practices to consumers so they can determine whether they were adequately served.

However, there is still unfinished business in bringing the instruction at APA polygraph schools into line with the same standards to which members are held. At this writing there is still no requirement for schools to teach the techniques members must use.

A few years ago a rather routine paper was published in *Polygraph* addressing the state of the research on polygraph techniques (Krapohl, 2006). The paper was nothing more than a summary of the published validity literature for the various polygraph techniques, complete with various tables and numbers and statistics and citations. Despite its dry and unassuming content, that paper ignited one of the greatest public debates in the field in decades. The contest pitted the literature summary against the opposing view of three prominent polygraph technique developers (see *Polygraph*, 2007, Vol 1) who had trained perhaps thousands of examiners in APA-approved schools. Their chief complaint: The article had not listed their favorite techniques as having been adequately researched.

One lesser known portion of the Krapohl (2006) paper covered the notion of "valid principles", that is, those individual practices that gave validity to polygraph techniques. Those principles were merely listed in the 2006 paper to make a point about how one might approach the development or evaluation of techniques. They were not fully supported or cited in the article, however, and I hope to remedy that shortcoming here. This monograph is a summary of 20 polygraph principles that have been supported by research. It is not an exhaustive list, nor is it the final word, but it is the collective evidence from various independent sources that can help practitioners identify and use the best polygraph methods. So, in no particular order, here are the:

Valid Principles for Polygraph Testing

1. There are no more than 12 reliable diagnostic tracing features in manual scoring. The three "Kircher features" may be sufficient alone.

Sources

Harris, Horner & McQuarrie (2000)
Kircher & Raskin (1988)
Kircher, Kristjansson, Gardner, & Webb (2004)

Background

The current and converging research findings on polygraph scoring point to simplicity in the analysis of the tracings. These findings agree quite well with the general conclusions in the field of diagnostics and decision theory. They are, however, in stark relief to long-held teachings in some portions of the polygraph profession where the number of scoring features can run into the dozens or embrace the even more ambiguous notion that "any change from the norm is a reaction." Those extreme views appear no longer tenable.

Though some level of skill is certainly necessary for the analysis of polygraph data, it would be an overstatement that the most complicated methods in the field lead to better accuracy. Much of the complexity of proprietary scoring systems appears to be unnecessary. For example, the three single most valid reactions for scoring, sometimes called "Kircher features", are respiration line length, electrodermal response amplitude and cardiovascular response amplitude. These responses are sufficiently powerful that it has been suggested that they could replace all other measures (Harris, Horner & McQuarrie, 2000; Kircher, Kristjansson, Gardner & Webb, 2004). They are also the features used for the CPS and OSS algorithms, both of which have met or exceeded the performance of experienced examiners conducting blind scoring (Kircher, Kristjansson, Gardner, & Webb, 2004; Nelson, Krapohl & Handler, 2008). The Kircher features are also at the core of the Empirical Scoring System (ESS), a simpler method which has shown a performance equivalent to or better than traditional methods (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Krapohl, 2010; Nelson, Krapohl & Handler, 2008).

The number of manual scoring features having empirical support is somewhat confused by how these features are described. For example, both respiration suppression and respiratory apnea are valid scoring features, but the latter is actually one form of the former. Similarly, the change in the inhalation/exhalation ratio typically co-occurs with respiration slowing, but these two features are usually denoted separately. All of the traditional features that are valid can be reduced to respiration line length (suppression, apnea, slowing) except the temporary rise in the baseline. The baseline rise is a strong feature, but it is not seen as frequently as the others. When it does occur, it is virtually always accompanied by suppression.

Electrodermal amplitude is the single most powerful predictor of the deceptive or truthful status of the examinee, accounting for about half of all of the diagnostic information available in the polygraph charts. Duration and complexity are weaker indicators, and not universally found in laboratory experiments.

In the cardiovascular channel, amplitude carries most of the weight, with duration providing additional information. Pulse constriction followed by pulse expansion is also a weak indicator.

The vasomotor channel relies on pulse constriction and duration. The best window for analysis is between 5 and 14 seconds after question onset (Cushman et al, in progress).

There may be circumstances where different features prove to be diagnostic, typically idiosyncratic patterns limited to single individuals. These features can be scored provided that the examiner can demonstrate the tracing feature is valid with the particular examinee. However, these features do not generalize to other examinees, and should not be used beyond those for which it is known to be a valid indicator.

2. Numerical scores tend to be more negative when an irrelevant question is placed immediately before a relevant question than when another evocative question (such as a probable lie) is just before the relevant question. This is true for both truthful and deceptive examinees.

Sources

Cullen & Bradley (2004)
Krapohl & Dutton (2005)

Background

In the early days of the field of polygraphy, examiners appeared less concerned about question sequencing,

sometimes making on-the-spot decisions about placement of each question. Perhaps the early pioneers were less familiar with orienting responses or the effect of the law of initial values, and how they can influence response magnitudes. Movement toward structured question sequences we now call "formats" developed later.

Cullen and Bradley (2004) were interested in the effect of question sequences on polygraph scores. They manipulated the sequences so that the relevant questions followed an irrelevant question, or the relevant questions followed a special type of "control question." What they found was that scores were significantly more negative in the former order than the latter. In fact, when the relevant question came immediately after an irrelevant question, the average score for truthful examinees was below zero. In a field test of the Cullen and Bradley hypothesis, Krapohl and Dutton (2005) used scores from the Federal Zone Comparison Technique (FZCT) cases and compared them to those of the Army Modified General Question Technique (AMGQT). The FZCT has a probable-lie comparison question placed immediately before each relevant question, whereas the first two relevant questions in the AMGQT were preceded by irrelevant questions. Consistent with the findings of Cullen and Bradley (2004), the FZCT data showed a relatively flat score profile across the three relevant questions, but the average score for the first two questions in the AMGQT was below zero for both truthtellers and liars.

3. A properly conducted Concealed Information Test (CIT) is as accurate as a properly conducted single-issue CQT.

Source
National Research Council meta-analysis (2003)

Background
We tend to think of the CIT narrowly as a particular approach in polygraphy. The CIT is actually more of a paradigm than an actual test. It is a template that can be laid over any situation where those who have crime-related information in their memories can be distinguished from those who do not have that information simply by exposing these individuals to relevant and relevant-like stimuli. The CIT works well in polygraphy using autonomic responses, but the CIT can also be used in tests that monitor body tremors, eye movements, brain waves, behavior, inter-personal spaces, word choices, and a variety of other measures.

The report of the National Research Council in 2003 suggests that the CIT in polygraphy, in particular using electrodermal and vasomotor responses, will have accuracy statistically similar to that of the CQT in event-specific, that is, specific-issue testing. Approaches for achieving maximum accuracy have subsequently been published (Meijer, Verschuere & Ben-Shakhar, 2011). While it has been argued that the CIT cannot be used nearly as often as the CQT (Podlesny, 1993; Podlesny, Nimmich & Budowle, 1995), and cannot be used in screening at all, nevertheless for circumstances where the CIT can be used it offers certain advantages over the CQT. The APA has published a how-to guide for the CIT (Krapohl, McCloughan & Senter, 2006).

4. On average, deceptive examinees react stronger to RQs than truthful examinees react to CQs.

Sources
Franz (1988)
Kircher & Raskin (1988)
Krapohl & McManus (1999)
Raskin, Kircher, Honts & Horowitz (1988)

Background
One of the long running assumptions in polygraphy is that examinees will react to either relevant or comparison questions, depending on whether they were being deceptive to the relevant questions. This assumption is manifested in the symmetrical cutoff scores used in the first seven-position scoring system (Backster, 1963), a method that was based on the untested hypothesis of balanced reactivity. The Backster method was amended in 1983 to include asymmetrical cutoffs (Backster, 1985), though there were no statistical analyses to support the new cutoffs. Shortly thereafter came a series of studies converging on the finding that the phenomenon in polygraphy was actually asymmetrical. For this reason, symmetrical cutting scores lain over an asymmetrical phenomenon would lead to unbalanced

accuracies. Because on average liars react to relevant questions more strongly than truth-tellers react to comparison questions, symmetrical cutoffs disadvantage truthful examinees. The asymmetry is exacerbated when the relevant questions are immediately preceded by irrelevant questions (See Principle 2).

Symmetrical cutoffs still prevail for the Utah Probable Lie Test (Bell, Raskin, Honts & Kircher, 1999), Horizontal Scoring System (Gordon, 1999; however, also see Nelson & Handler, 2012; Krapohl, Gordon & Lombardi, 2008) and the Federal ZCT (Light, 1999), though the Federal ZCT also employs the Spot Score Rule that further shifts the emphasis toward the detection of deception. Exceptions to the trend in symmetry include the Matte system (1999) and the Empirical Scoring System (Blalock, Cushman & Nelson, 2009; Nelson, Krapohl & Handler, 2008) both of which established cutoffs from evaluation of normative field data.

It is not a fair statement that asymmetrical cutoffs are always best until one operationally defines "best". They can lead to balanced accuracy, true, but because all tests have errors one must also consider whether the costs of errors are also balanced. Expressed practically, is the cost of missing a liar always greater than that from missing a truthteller? Frequently false positive and false negative errors have very different costs, and can vary by context. For example, in criminal testing there is a higher cost for false negatives (letting a liar go) than false positives (additional questioning). This imbalance between the two types of errors may explain why many agencies adhere to risk-aversive cutoffs that are so much better at detecting lies than truthfulness (Blackwell, 1999). These cutoffs typically have evolved from the practices of field examiners without the benefit of statistical assessment. A more rational approach can be found in decision theory, where cutoffs are established according to the likelihood of, and tolerance for certain errors. However, with the exception of the Empirical Scoring System and the Objective Scoring System, there is currently no way to calculate decision errors from manual scores, an essential component for assessing the error likelihoods at specific cutoff scores. Moreover, no manual scoring system yet published sets cutoff scores according to a cost-benefit analysis. Much work remains to be done in this area.

5. Countermeasure sensors improve detection of physical countermeasures.

Sources
Honts, Raskin & Kircher (1983)
Ogilvie & Dutton (2008)

Background

Countermeasures seemed to have evolved right along with the lie detection field (Benussi, 1914; Reid, 1945; Stewart, 1941). In 1945, John Reid developed an examination chair with sensors to detect covert movements, a forerunner of today's approach to countermeasure detection.

There is ample evidence that physical countermeasures can dramatically reduce the detection of deception in the absence of sensors (Honts, 1984; Honts & Hodes, 1983; Honts, Hodes & Raskin, 1985; Honts, Raskin & Kircher, 1983). For this reason, the American Polygraph Association has mandated the use of countermeasure sensors for all testing beginning in 2012, and the US government has had the requirement since at least 2006 (Federal PDD Examiner's Handbook). These sensors have not proven useful for other types of countermeasures, but are considered essential in detecting and deterring the one type shown to be effective: physical countermeasures.

6. Exams that start with a demonstration test have better accuracy than those that do not.

Sources
Gustafson & Orne (1965)
Kircher, Packer, Bell & Bernhardt (2001)

Background

The demonstration test (AKA stim test, acquaintance test) has been used for decades, and promoted by almost all schools of thought. These tests serve several purposes, and among them are: familiarizing the examinee with sensors and procedures; to gather physiological data that might be useful in determining the examinee's norm or for evidence of countermeasures, and; to ensure

the sensors are properly placed and the gain settings are correct. Evidence has shown that conducting the demonstration test can improve polygraph decision accuracy, especially if feedback is given to the examinee that the demonstration test worked well with him. There is insufficient evidence to point to any particular form of demonstration test as the best.

7. Conventional scoring methods of directed-lie data are not effective in the pneumograph

Sources
Horowitz, Kircher, Honts, & Raskin (1997)
Kircher, Packard, Bell, & Bernhardt (2001)
Dollins, Pollina, & Krapohl (NCCA02-R-0004, unpublished)

Background

The very first polygraph screening technique to be developed and validated by scientists was the Test for Espionage and Sabotage (TES; Research Division Staff, 1995a, 1995b). Among TES's innovations was the inclusion of repeated directed-lie comparison questions in the series. Directed lies had been in use since at least the late 1960s (Fuse, 1982; Menges, 2004), and field practitioners had come to recognize the difference response patterns in the pneumograph as compared to probable-lie comparison (PLC) questions. In the initial TES research, however, examiners made no distinction between the features of the PLC and those of the DLC. Because there had been no effort to explore the contribution of each polygraph channel toward decision accuracy, the effect of using PLC scoring features on DLC data in the pneumograph remained undiscovered until the 1990s. It is now generally accepted among both examiners and researchers that conventional methods of scoring the pneumograph are not valid when using DLC polygraph techniques.

8. Two-stage (Senter) rules can reduce INCs without affecting total decision accuracy.

Sources
Blalock, Cushman & Nelson (2009)
Krapohl (2005)
Krapohl & Cushman (2006)
Nelson, Krapohl & Handler (2008)
Senter (2003)
Senter & Dollins (2008)

Background

In an innovative departure from the standard approach to polygraph decision rules, Senter (2003) found that employing more than one step could deliver high accuracy at a significantly lower Inconclusive rate. It begins by the use of the total score only. If the total score exceeds a cutoff, an NDI decision is made, irrespective of any subtotal score (aka spot score). If the total is lower than the other cutoff, a DI decision is made. These cutoffs will depend on whether the standard federal preference is used (+/-6), evidentiary decision rules (+4, -6), or the standard decision rules of the Empirical Scoring System[2] (+2, -4). Only when the total score falls between these cutoffs is the second stage engaged. The second stage entails the use of the subtotal scores, or the totals of each individual relevant question. When the subtotal score is lower than the cutoff, a DI decision is made. Again, the subtotal score cutoff will depend on the scoring system: -3 or lower for the federal and evidentiary decision rules, and -7 for ESS. The research evidence points to an average reduction of about 60% in inconclusive results and no loss of decision accuracy when the Senter Rules are part of the decision rules.

9. Polygraph decision accuracy is not associated with the gender of the examinee.

Sources
Reed (1993a)
Buckley & Senese (1991)
Kircher, Packard, Bell & Berhardt (2001)

Background

There are well established gender differences in psychophysiological tonic and phasic behavior (for a review, see Anderson & McNeilly, 1991). One may be tempted to conclude that these differences would manifest themselves in polygraph decision

[2] Because the ESS cutoffs are based on normative data, they can be adjusted such that the examiner can reduce the probability of either false positive or false negative error rates to levels that correspond with the cost of those errors.

accuracy. Indeed, polygraph research examining gender differences have intermittently found a tendency for females and males to respond in different polygraph data channels (Bradley & Cullen, 1993; Matte & Reuss, 1992, but also see Miyake, 1978). However, a connection between polygraph decision accuracy and examinee gender has not been reported. One reason that significant differences in veracity decisions for male and female examinees would be unexpected is that conventional polygraphy uses ipsative (within-subject) analyses in that the examinee's responses are compared to other responses that same examinee produces. Examinee responses are not compared to those of other examinees where gender might affect the interpretation. Also, conventional polygraphy uses multiple physiological channels of data, and decisions are based on aggregate scores across all channels. Individual or gender factors may affect which physiological channel provides the most useful information, but those differences are lost when aggregate scores are used for decision making. Consequently, gross measures such as polygraph decisions would not be expected to be sensitive to gender.

10. Polygraph decision accuracy is not different between African American and Caucasian examinees.

Sources
Buckley & Senese (1991)
Reed (1993b)
Krapohl & Gary (2004)

Background

As with gender, there is a long and rich body of research pointing to racial differences in physiological responding between Caucasians and African Americans (Anderson & McNeilly, 1991). Whether these differences manifest themselves in polygraph data becomes an important question. One of the most prominent figures in the polygraph field has long contended that the cardiovascular channel provides a better indicator of veracity among African American examinees than it does for Caucasian examinees (Arther, 1998). If there are differences in the patterns of responses that correspond with racial groups, it is theoretically possible to tailor scoring and algorithmic systems to those groups to improve decision accuracy.

To date, no data-derived scoring system has been published that attempts to accommodate differences in response patterns from demographic groups. With what is known about the wide inter-examinee variability in physiological responding, it appears unlikely such an endeavor would be productive. Early evidence of racial factors producing different profiles of response patterns has been disappointing (Krapohl & Gary, 2004). What has been demonstrated, however, is that whatever underlying differences there may be they have not been shown to influence polygraph decision accuracy.

11. The only technical question found to improve polygraph decision accuracy is the comparison question.

Source
Cushman & Krapohl summary of multiple studies (2010)

Background

Even the earliest pioneers of the polygraph field recognized that accuracy of their techniques fell below perfection. With the aim of improving accuracy, many early practitioners used a trial-and-error approach, conducting informal field experiments on live cases while tracking successes. Others based their methods on hypotheses about what caused polygraph errors. Formal methodology and statistics have long been crude, incomplete or absent among examiner/ researchers. Virtually none of the evaluations of these methods would meet current understandings of systematic investigation.

One persistent line of thinking for boosting accuracy has been that the examiner could add other kinds of questions to the test (generically called "technical questions" here) that could: provide a benchmark against which to evaluate the charts; mitigate errors by identifying outside factors that compromise accuracy, or; reassure innocent examinees for the purpose of reducing their reactions to relevant questions. How well has this approach worked? Table 1 lists the current state of understanding regarding the most common technical questions. With the exception of comparison questions, the research has generally not been supportive of technical questions. The trend of research

Table 1. Summary of validity research by technical question. (From Cushman & Krapohl, 2010. Used with permission)

Technical Question	Published Research?	Supportive Research?
Countermeasure	**Yes**	**No**
Sacrifice Relevant	**Yes**	**No**
Symptomatic	**Yes**	**No**
Hope/Fear	**Yes**	**?[1]**
Exclusionary Comparison	**Yes**	**Yes[2]**
Inclusionary Comparison	**Yes**	**Yes**
Directed-Lie	**Yes**	**Yes**
Positive Control	**Yes**	**Yes[3]**
False Key	**No**	**NA**
Known Truth	**No**	**NA**
Situational Control	**No**	**NA**
SKY	**No**	**NA**
Guilt Complex	**No**	**NA**

1. Not tested for the factor for which it was designed.

2. Did not perform as well as the Inclusionary Comparison Question.

3. Worked best when combined with Probable-Lie Comparison Questions.

might be interpreted as suggesting an end to the search for technical fixes for polygraph testing imperfections.

12. Non-exclusive (or inclusive) probable-lie comparison questions perform equal to, or better than, exclusive probable-lie comparison questions.

Sources
Amsel (1999)
Horvath (1988)
Horvath & Palmatier (2008)
Podlesny & Raskin (1978)

Background

One of the purported improvements to polygraph techniques was the introduction of the exclusive probable-lie comparison question. Exclusive comparison questions use devices such as time or place bars within the question to delineate them from the relevant question (e.g., Before the age of 23 did you ever steal anything?) The hypothesis is that a comparison question with any degree of overlap with the relevant issue could be confusing to the examinee, and cause some amount of reactivity from deceptive examinees to be expressed on the comparison questions, resulting in an increased chance of decision error. If this hypothesis is true, there would be a higher incidence of false negatives when the non-exclusive comparison questions are used versus the exclusive comparison questions. The remedy is to clearly differentiate the relevant and comparison questions from one another using delimiting language in the comparison questions. This hypothesis came to be "common sense" to most polygraph examiners long before it had been experimentally tested.

An alternate view is that narrower comparison questions (i.e., those with exclusionary phrases) have a lesser power to evoke reactions among truthful examinees. It is argued that comparison questions that are broader and more ambiguous are more likely

to generate deception or uncertainty to the comparison questions and make them more effective. If this hypothesis is true, one could predict that false positive results would be more common for exclusive comparison questions than for non-exclusive comparison questions.

Currently there are four published studies on which to assess the relative value of exclusive and non-exclusive questions: Three are laboratory studies, and the fourth used field data. The following is a summary of their conclusions:

Amsel (1999) was an examination of scores produced by exclusive and non-exclusive comparison questions in field cases. With a sample of 230 cases Amsel found that non-exclusive comparison questions tended to have stronger average scores in the correct direction for both truthful and deceptive examinees than those cases where exclusive comparison questions were used. This finding would be consistent with those that support the non-exclusive comparison question, and not support the use of time or place bars. Amsel's findings were subsequently criticized by Matte and Backster (2000) for: using three charts per case; use of 3-position scoring; violating the Backster concept of the sacrifice relevant question; and having a comparison question as the last question on the test. These factors had been held constant for both the exclusive and non-exclusive comparison question in the Amsel field study to isolate any significant findings to the questions themselves.

In a laboratory study using a Modified General Question Test (MGQT), Horvath (1988) tested 60 volunteers that had been programmed as either guilty or innocent of stealing cash from an office. As with the Amsel study, blind scoring of Horvath's data pointed to an advantage to using the non-exclusive comparison question. In a direct test of the hypothesis that guilty examinees may confuse non-exclusive comparison questions with relevant questions, and thereby diminish their scores, Horvath found quite the opposite. Scores of deceptive examinees were actually stronger with non-exclusive comparison questions. Overall, non-exclusive comparison questions reduced decision errors.

One of the possible criticisms of the Horvath (1988) study was that the comparison questions had not been assessed in a Zone Comparison Technique (ZCT) format, limiting his findings to a technique not used as often in the field. To address this criticism, Horvath and Palmatier (2008) looked at the two types of comparison questions in both the MGQT and the ZCT in another laboratory study. Again, in this head-to-head analysis, non-exclusive comparison questions outperformed the exclusive comparison questions. Non-exclusive comparison questions produced significantly higher accuracy and fewer false positives. In no measure of effectiveness did the exclusive comparison question surpass the non-exclusive comparison question.

In the first study to compare the two types of comparison question, Podlesny and Raskin (1978) found no significant differences in decision accuracy between exclusive and non-exclusive comparison questions. Their data did suggest that certain physiological responses were more discriminative for the exclusive comparison question: mean skin conductance response recovery half-time, mean skin conductance recovery half-time width, and mean negative skin potential response amplitude. Because none of these features are used in any conventional scoring system, their practical value would be negligible.

In sum, the existing studies do not support the "common sense" hypothesis that exclusive comparison questions offer advantages in terms of decision accuracy. The shared finding among all of the available research is that non-exclusive comparison questions do as well as, or better than, exclusive comparison questions.

13. The Utah 3-to-5 chart rule can significantly improve decision accuracy.

Sources
Senter & Dollins (2004)
Senter, Dollins & Krapohl (2004)

Background

By way of explanation, the 3-to-5 chart rule specifies that when the scores of three charts would lead to an inconclusive call, two more charts are collected and scored. The scores from all five charts are added together

for the final decision, which is based on the same cutoffs as the three-chart cutoffs.

The Utah Probable-Lie Test began incorporating the 3-to-5 chart rule fairly early in the development of that technique. Whether, and how much, the final two charts improved efficacy was not investigated until the work of Senter and collaborators in recent years. Senter, Dollins and Krapohl (2004) were the first to find a significant improvement in decision accuracy when the 3-to-5 chart rule was used as compared to analyzing only three charts. In the subsequent Senter and Dollins replication (2004), the rule boosted decision accuracy about 8%, another significant finding. Senter and his collaborators determined that differences in decision accuracy between the Utah scores and those from the federal government were not attributable to how scores are assigned by the scorers. Rather, their data suggested the better accuracy of the Utah system could be isolated to three factors: the 3-to-5 chart rule, the addition of the photoplethysmograph, and setting aside the Spot Score Rule (Light, 1999) in favor of using only the total score. The greatest contribution came from the additional data. In a related finding, Senter (2003) determined that inconclusives could be suppressed without affecting decision accuracy by employing two-stage decision rules (see Principle 8).

14. The Friendly Polygraph Examiner Hypothesis (Orne, 1973) is not valid for the comparison question technique.

Sources
Honts (1997)
Matte & Reuss (1990)
Orne (1973)
Raskin (1976)

Background

Martin Orne (1973) proposed that polygraph examinations conducted under conditions where there were no adverse consequences for failing would tend to produce false negative results. The premise is based on the idea that the physiological reactions that are essential to diagnosing deception are generated by the fear of detection, or at least fear of some punishment if the deception were detected. Absent this fear, such as testing conducted under confidentiality of one's defense attorney, Orne's expectation was that guilty examinees would not produce the requisite physiological reactions and thereby go undetected.

Orne's hypothesis appeared attractive among critics of polygraphy because it suggested that polygraph accuracy could vary according to who conducted the test. The hypothesis was also embraced by some polygraph examiners working in the criminal justice system who were suspicious of privately conducted polygraph testing on behalf of a defendant. The testimony of members of both groups convinced some courts as to the validity of it.

Even on its face, though, Orne's Friendly Polygraph Examiner Hypothesis (FPEH) contains logical errors. As Honts (1997) observed, the FPEH assumes that the underlying cause of polygraph reactions is fear, and that there is no fear when polygraph examinations are conducted under defense attorney privilege. Consequently, the FPEH leads to the conclusion that confidential polygraph examinations would be vulnerable to false negative results. In contrast to the first assumption regarding the necessity of fear in polygraphy, decision accuracy has been demonstrated in many laboratory settings where the level of fear is far less than field conditions, perhaps even absent. Consequently, the data suggest fear may not be necessary for the polygraph technique to be effective (See Handler, Shaw & Gougler, 2010; Khan, Nelson & Handler, 2009). It is also important to recall that virtually all polygraph examinations conducted in modern times on criminal matters are a variation of the comparison question technique (CQT). For the CQT to produce anything but inconclusive results, examinees must react to either the relevant or comparison questions. It is the differential salience between these two categories of questions that give rise to the differential reactivity on which veracity decisions are based (Handler & Nelson, 2007; Senter, Weatherman, Krapohl & Horvath, 2010). A lack of reactivity, as supposed in the FPEH, would not lead to a false negative decision but an inconclusive outcome.

On the evidence side, Honts (1997) added that in his own polygraph practice there was no significant difference in the

proportions of deceptive decisions for confidential and non-confidential examinations. Similar findings were reported by Matte and Reuss (1990) and Raskin (1976). There is no published report supporting the FPEH for the comparison question technique, and given the contrary field evidence, it does not appear to be true.

15. For police screening, the polygraph topics most predictive of officer success address criminal behavior (including drugs), disciplinary action from previous employers, and tolerance (domestic violence, racial/ethnic slurs directed at individuals).

Sources
Aamodt (2004)
Handler, Honts, Krapohl, Nelson, & Griffin (2009)

Background

The selection of polygraph test topics in police screening almost always has one of two originators: department leadership or polygraph examiners. Department leaders have the final responsibility to select the best candidates, and they use criteria they believe predictive of job success in their screening process. In many cases, they direct polygraph examiners on the topics they will use in the screening police candidates. Examiners have considerable training and expertise in question development appropriate for polygraph testing, and rightly have the ultimate responsibility in crafting and refining the questions used in their examinations. If they do not receive guidance from the department superiors, examiners typically will turn to information supplied by other examiners or use their own experience to decide which questions should be covered in screening examinations.

Further up the decision chain should be, but rarely is, the empirical support on which to base the selection of topics used in polygraph screening examinations. Though much research has been published, very few departments avail themselves of it but rely instead on their own best judgment. That judgment varies from department to department, and even among leadership within a department. Consequently, there is no standardization across police polygraph screening programs, resulting in an immense variety of topics among police departments. Testing conducted on behalf of Department A may cover topics with little overlap to polygraph testing conducted for Department B, though they might both be seeking the very same type of candidate. The departments may also be using polygraph topics with questionable or little nexus to the job responsibilities, leading to the selection or non-selection of applicants based on factors with no predictive value. These conditions compromise the potential efficiency, effectiveness and validity of the polygraph process, to say nothing about additive costs to the department or fairness to the candidates.

Michael Aamodt (2004) has summarized the research on police candidate selection, and from his work polygraph examiners can determine which factors are amenable to polygraph testing (see Handler, Honts, Krapohl, Nelson & Griffin, 2009). They are:

1. Criminal behavior (including drugs)
2. Past disciplinary action by employers
3. Tolerance (e.g. domestic violence, racial and ethnic slurs against individuals, history of excessive force)

These topics might be covered differently by individual departments (i.e., dividing them among three to eight individual polygraph test questions) so long as each topic is thoroughly tested. Departments also frequently choose to ask about the accuracy of the candidate's applicant documents during the polygraph examination. Though relevant information can be gleaned when this topic is included among the relevant issues, it is also time-consuming, marginally productive, and the documents are already normally verified by the routine background investigations.

There may be other areas of unique interest to certain agencies due to patterns of problems observed in the workforce, or because of penetration attempts by gangs, organized crime, or foreign governments. These questions should also be included in the questions list as appropriate. Examiners should work with hiring officials to identify those behaviors that can be tested by the polygraph and also be justified as being predictors of future problems. The use of the polygraph as a "fishing expedition" has been

the source of many public relations problems for polygraphy for decades, and has been one significant contributor to legislation that has restricted the field.

16. The 3-position Empirical Scoring System performs at least as well as the traditional 7-position scoring system.

Sources
Blalock, Cushman & Nelson (2009)
Handler, Nelson, Goodson & Hicks (2010)
Nelson, Blalock, Cushman & Oelrich (2011)
Nelson, Handler, Shaw, Gougler, Blalock, Russell, Cushman & Oelrich (2011)
Nelson & Krapohl (2011)
Nelson, Krapohl & Handler (2008)

Background

Unlike most polygraph scoring systems in common practice, the Empirical Scoring System (ESS) was not simply a mutation of an earlier system. Rather, it began as a zero-based reviewed of the psychophysiological and decision theory literature. Establishing their approach on a significant body of scientific findings, the developers then went on to "assemble" more than "invent" the ESS. They began with the identification of which scoring features produced the most diagnostic information. Based on the principle of parsimony, they simplified the system to the degree possible, and then built a database of normative data on which to establish cutoff scores. This step-wise methodology - the building of each procedure on a firm scientific footing - gave the ESS something quite valuable: the traceability of the final decision through demonstrably defensible steps. And with it, the departure from the historic "faith-based" scoring systems.

Theory is one thing, and however beneficial that scientific support might be, the ESS would have little practical value if it did not also produce acceptable accuracy. Over the course of the past few years the developers and independent researchers have applied the ESS to several samples of confirmed cases. What has been shown is that ESS decision accuracy is always equal to, or better than that from scorers using other traditional methods. Inter-scorer agreement is also high, an expected conclusion given the simplicity of the ESS.

One criticism of the ESS is that it has a lower sensitivity than do other scoring systems. The prospect that the ESS may mis-call more deceptive cases than another system is one rationale for some examiners remaining with the traditional scoring systems. This justification has overlooked an important characteristic of the ESS. Because the ESS derives its cutoffs from normative data, it is possible to select ESS cutoffs that match the customer's tolerance for false negatives and positives as well as inconclusives. If a customer is risk-aversive, the polygraph examiner can select cutoff scores that minimize false negative errors. In applications where false positives have hefty consequences, such as in evidentiary applications, cutoffs can be established that have a balance of false negatives and positives. Not only does the ESS allow examiners to render the traditional NDI, DI, and Inconclusive decisions, but also to report a probability of error of these decisions.

ESS is currently taught at some polygraph schools, and is gaining acceptance. For those interested in learning more about the ESS, see Nelson, Handler, Shaw, Gougler, Blalock, Russell, Cushman and Oelrich (2011).

17. Inter-chart discussions reduce decision errors.

Sources
Dawson (1981)
Honts (1999) Research summary.
Honts, Hodes & Raskin (1985)
Honts, Raskin & Kircher (1987)
Honts, Raskin & Kircher (1994)
Horowitz, Raskin, Honts & Kircher (1997)
Kircher & Raskin (1988)
Patrick & Iacono (1989)
Podlesny & Raskin (1978)
Raskin & Hare (1978)

Background

Habituation across charts has been shown to be a factor in comparison question testing (Kircher, Raskin & Honts, 1984; Stern & Kircher, 2002). One approach to dishabituate examinees is for the examiner to briefly discuss the relevant and comparison questions between charts. The goal is to help maintain a level of arousal throughout testing, and to ensure questions remain salient to the

examinee. This practice, however, has met resistance from the examiner community (Abrams, 1999; Matte, 2000). There is a concern that inadvertent (or not) emphasis on one category of question over another can tip the response pattern of the examinee in one direction or another.

Taken to an extreme, an adverse effect for unbalanced emphasis on certain questions is probably unarguable. Differential salience (Senter, Weatherman, Krapohl & Horvath, 2010) can most certainly be manipulated by the conditioning done by the examiner, both before the test and during the test. However, the other extreme, that generic or balanced discussions between charts are harmful, is not so tenable. Merely reviewing the questions between charts and asking the examinee whether he is still comfortable with his answers has no obvious drawbacks, and may help avoid the problem of increasingly flat charts. Moreover, prohibiting the practice of inter-chart discussions, as many polygraph schools do, raises the obvious question: if discussion of the questions immediately before the first chart is acceptable, why should it be prohibited before the other charts?

As a general operating principle it may be important to recall that habituation will be a problem in some exams, and that measures such as inter-chart discussions can serve to keep the examinee engaged. Other tools are also available, of course, such as having the examinee repeat a keyword from each question with his answer, and interspersing *yes* answer and *no* answer questions in the sequence to ensure the examinee's attention. Each of these methods can be helpful when used judiciously.

18. The Spot Score Rule does not improve decision accuracy for specific issue testing.

Sources
Hedges & Deitchman (2012)
Senter & Dollins (2004)
Senter, Dollins & Krapohl (2004)

Background

There are considerable differences among the various polygraph scoring systems regarding the decision rules for rendering DI and NDI results: per-chart minima, per-question totals, whole examination totals, etc. One fairly common decision rule is the Spot Score Rule (SSR; Light, 1999).

As previously outlined in this article, the SSR can trump the total score in the decision process by forcing a DI call when the spot score is -3 or lower, irrespective of the grand total score. Without question, the addition of the SSR improves the sensitivity of the polygraph test to detect deception, but does not improve the accuracy of the test as a whole. The improvement in the detection of deception comes at the reduction of the ability of the test to detect truthfulness. More about this later.

There are two factors working synergistically to reduce the detection of truthfulness. One is variability. It is commonly accepted that smaller samples are more variable than larger ones. A spot score represents only a minority of all of the scores in an examination, and consequently it will vary more proportionately than will the total score. This variability virtually ensures that some percentage of individual spot scores will fall below the spot score threshold than the more stable total score for the grand total threshold. For example, consider the data from the Blackwell (1999) study where three federal scorers evaluated 35 truthful and 65 deceptive field cases.

In a re-analysis of the 65 deceptive cases in the Blackwell study, an average of 46.3 (71%) would have been DI by the -6 total score threshold. Another 11.7 (17%) were correctly classified by the SSR. A total of 2% of the deceptive cases would have been called NDI by the total score of +6 or greater had not the SSR not intervened. See Table 2.

Looking at the net effect, the correct classification of deceptive cases improved 17% when the SSR was added over using the total score to base the decisions. Truthful cases experienced a 19% decrement in accuracy when the SSR was used instead of the total score. Using the total score, there would be about a 2% false negative rate, but adding the SSR requiring all spots to have positive values reduced this to 1%. Using the total score only, average decision error was 3.9% (average of 5.7% and 2.1%). The SSR -3 rule alone resulted in a false positive rate of nearly 25%.

Table 2. Average effect of spot scores on decision accuracy of 35 truthful and 65 deceptive cases for three federal scorers who participated in the Blackwell (1999) study. In percent.

	Truthful Cases	Deceptive Cases
All spot scores >0	44.8	1.0
Any spot score <1	53.3	99.0
Any spot score <-2	24.8	89.2
DI decision based on total score	5.7	71.3
NDI decision based on total score	63.8	2.1
Inconclusive based on total score	30.5	26.7

In this data set, the SSR reduced decision accuracy overall.

As one can see from this re-examination of the Blackwell (1999) data, the SSR did not improve the accuracy of the test. It increased total error to a substantial degree. The reason for this will be taken up later in this section

To test whether this effect was restricted to the Blackwell sample, a separate analysis was conducted using scores produced by five federal polygraph examiners who blind scored the cases previously used by Krapohl and Cushman (2006). Looking again at the net effect, there was a similar finding as with the Blackwell (1999) cases. See Table 3. The SSR boosted detection of liars, but incurred a significant loss in detection of truthful cases. Average decision error using only the total score was 3.4% (average of 2% and 4.8%) to an error rate of 20% for truthful cases alone. This trend indicated that the SSR reduced decision accuracy, consistent with the Blackwell data.

Table 3. Average effect of spot scores on decision accuracy of 50 truthful and 50 deceptive cases for five federal scorers. In percent.

	Truthful Cases	Deceptive Cases
All spots scores >0	50.4	3.6
Any spot score <1	49.6	96.4
Any spot score <-2	20.0	82.0
DI decision based on total score	2.0	55.6
NDI decision based on total score	62.0	4.8
Inconclusive based on total score	36.0	39.6

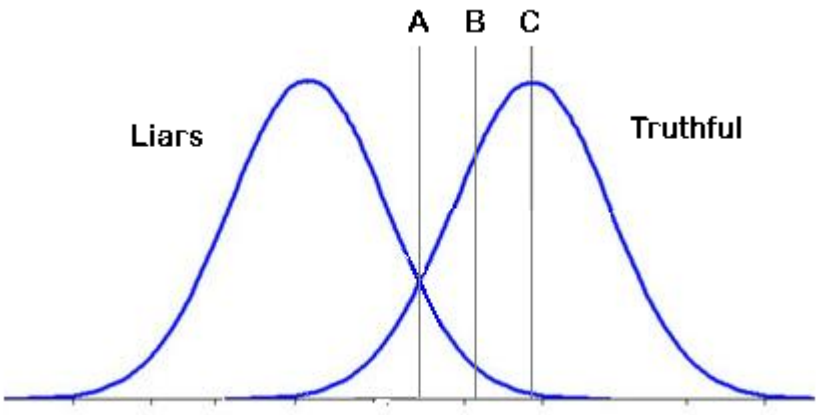
Why the SSR is perceived by some as an improvement in decision accuracy may be attributable to a perception influenced by context. In settings where the base rate of deception is high, such as the many polygraph programs that test mostly prime suspects, the use of the SSR allows examiners to make a higher proportion of correct decisions simply because the SSR is good at detecting what these examiners are facing the most: liars. Think of it like this: if a DI call were to be made when a flipped coin landed on "heads", and virtually all of the examinees are actually liars, errors would be minimized if one used a two-headed coin. If the polygraph were used in a low base rate setting, it is likely that the

lopsided performance of the SSR would certainly be more noticeable.

There is an additional problem with the SSR beyond the wide variability of spots scores that often throws truthful cases into the deceptive category. It is the SSR's exceptional steps to avoid false negatives. It is axiomatic that decision rules that move false negative error rates closer and closer to zero will increase false positive errors, not in proportion, but incrementally faster than the reduction in false negatives. The effect is best explained conceptually using the familiar bell curves.

As with most measurements of human characteristics, the frequency of polygraph scores tends to fall into two overlapping distributions (See Figure 1.) For sake of illustration, the two overlapping bell curves in Figure 1 are meant to represent the frequency of polygraph scores on which decisions are based. For simplicity, this thought exercise will only consider total scores, but the same principle would apply to most decision rules, including the SSR.

Figure 1. Bell curves representing a hypothetical frequency distributions of scores for liars and truthtellers, along with three possible cutoff points.



Let us, for the moment, agree that the curve on the left represents the distribution of scores for liars, and the other bell curve is the distribution of scores for truthtellers. As all examiners know, most liars tend to have total scores below 0 while the opposite is true for truthtellers. The lines marked A, B and C are hypothetical cutoff scores, which we will consider separately.

If the cutoff score used to make a DI or NDI decision (ignoring for the moment inconclusive calls) were placed at the line marked "A", all of the scores to the left would be called DI, and all of the scores to the right would be called NDI. The portions of the curve on the wrong side of the cutoffs would be errors. For liars, the error rate would be the portion the liars' curve to the right of cutoff A, and for truthtellers the errors are those cases falling to the left of A. At cutoff A, there is a balance in accuracy: there is an equal proportion of errors for both the truthteller group and the liar group. At this point decision errors overall are at their lowest. This fact may explain why most algorithms choose decision points that afford balanced accuracy.

Suppose the user did not want to make as many false negative errors (miss liars). Going then to cutoff B would capture more liars than cutoff A. Notice that the line cuts more to the far end of the liars' bell curve, meaning that more liars are falling below this cutoff than would at cutoff A and thereby catching more of the liars. However, cutoff B would also misclassify a new proportion of truthtellers, and in a larger proportion than the improvement in detecting

liars. This is the tradeoff, catch a few more liars but miss more truthtellers.

Suppose now that one wanted to avoid false negative errors almost completely. In that case, cutoff C would be the best choice. Observe that going to cutoff C would allow detection of virtually all liars. There are hardly any liars with scores to the right of cutoff C. This same cutoff would misclassify approximately half of the truthtellers, however. This is the lesson of this thought exercise. The loss and gain becomes increasingly unequal as the cutoffs try to avoid one type of error. At extreme levels, detection one group can be at chance levels, or even below chance levels. The prevailing blind scoring results with the Federal ZCT shows that it has a remarkable capacity to detect liars and makes virtually no false negative errors. However, its detection of truthtellers hovers in the range of 45% - 60%, suggesting that the users of the SSR are exceptionally concerned about false negative errors as compared to false positive ones. The lopsided performance is predicted, and explained, by Figure 1.

This exercise is not meant to suggest that the SSR is undesirable in all situations: far from it. When the cost of a false negative error is great (e.g., missing a possible terrorist or presidential assassin) the SSR is easily justified. This is even more true if the consequences for a false positive are relatively trivial (e.g., an interrogation that would have happened even without the polygraph examination), or when decisions regarding actions against the examinee consider other sources of information or additional testing. The SSR may be the best choice in certain circumstances, but it would be a mistake to apply it to all contexts or to suggest the SSR improves overall polygraph decision accuracy. Both the theory, and the evidence, indicates this to be untrue.

19. Algorithms use diagnostic information more efficiently than do most human blind scorers.

Sources

Kircher, Kristjansson, Gardner, & Webb (2005).
Kircher & Raskin (1988)
Krapohl & McManus (1999)
Krapohl & Norris (2000)
Nelson & Handler (2012)
Nelson, Krapohl & Handler (2008)
Podlesny & Kircher (1999)

Background

Manual scoring of polygraph data is little more than an accounting system. While global interpretation attempts to reach decisions by overall impressions of the charts, manual scoring entails the assignment of numbers to very small subsets of the data, and then the tallying of the numbers at the end. The advantage of numerical scoring systems, at least the very good ones, is that they properly weight the significance and frequency of physiological events in a way that not only leads to valid conclusions, but provides a framework for other scorers similarly trained to come to the same conclusions. Global analysis can deliver accurate results in many cases, but the emphasis on the "art" and the lack of objective quantification invites more disagreement among scorers. It is generally recognized that increased disagreement translates into decreased accuracy, and the findings from studies comparing global and numerical analysis systems have generally conformed to the expectation that global does not perform as well as valid numerical scoring systems (Crowe, Chimarys & Schwartz, 1988; Ginton, Daie, Elaad & Ben-Shakhar, 1982).

Because scoring may improve both validity and reliability, automatic quantification through algorithms has long been of interest in the field of polygraphy (See Peters, 2011). Currently there are several decision algorithms available on computer polygraphs. For those algorithms that have been compared against human scoring, the algorithms tend to prevail. (Kircher, Kristjansson, Gardner & Webb, 2005; Kircher & Raskin, 1998; Krapohl & McManus, 1999; Krapohl & Norris, 2000; Nelson & Handler, in press; Nelson, Krapohl & Handler, 2008; Podlesny & Kircher, 1999). Some individual scores can, and do, outperform the algorithms in these studies; however, the striking majority of scorers do not. The accuracy of the algorithms is more impressive when considering that studies typically use very experienced or specially selected examiners as the manual scorers. And even for those individual scorers who bested the algorithm,

an open question is whether they can do so repeatedly.

There are no unmixed blessings in polygraphy, and it is important to note that in their current stage of development algorithms are not very efficient in the detection of artifacts, countermeasures, and anomalies. Consequently, experienced humans still play a critical role in the algorithmic process. Another consideration is that algorithms have set cutoffs for decisions of DI, NDI and Inconclusive that may not match the needs of the user. Most algorithm developers want to maximize decision accuracy, and this objective is furthered by using decision rules that produce balanced accuracy (see the previous section on the Spot Score Rule). Most algorithms have cutoffs that produce roughly equal proportions of false positives and false negative errors. While balanced accuracy tends to produce the highest accuracy, it also assumes the user agrees with the proportions of errors that balanced accuracy renders. This assumption is frequently false, inasmuch as investigative polygraphy values true positives and strenuously avoids false negative errors. Consequently, traditional scoring and decision rules match the values of investigative polygraph examiners more often than do algorithms, despite the lower overall accuracy of traditional scoring and decision rules. Moreover, the "successive hurdles" approach now embraced by many examiners can mitigate the false positive problem that arises in traditional manual scoring. Therefore, with extra effort and special practices, most polygraph examiners should be able to deliver competitive accuracy with standard scoring methods.

20. Those who research their own lie detection techniques, or use their own field cases, report accuracies at or near perfection.

Sources
See Table 4.

Background

The issue of polygraph decision accuracy has important implications for law enforcement, governments, examiners and examinees. This is because of what polygraph results can do. Polygraph results affect the lives, reputations, opportunities, and liberties of examinees, and often their loved ones, and just how much one can take action on polygraph results depends on how accurate they are.

The question of accuracy was thoroughly addressed in the National Research Council's report of 2003. Based on all available evidence, they placed the median percentage in the upper 80s for event-specific polygraph examinations. In the polygraph community, however, there is a more generous assessment. Official APA statements would have the percentage in the mid to high 90s, or even higher. How can two organizations looking at the same body of evidence come to such a divergent viewpoint?

The NRC (2003) had specific *a priori* criteria for acceptance of studies for their analyses which were included in the report. The polygraph community has been inclined to a more liberal interpretation of what constitutes evidence. Example: Prior to 2012 the APA considered a technique valid if it had published research, or (and this is important) was taught at an APA accredited school. In other words, school directors could decide which techniques would be considered valid irrespective of the existence of empirical support. This metric for validity is peculiar to polygraphy, and a comparable validity criterion in another legitimate field would be difficult to find. Fortunately, the APA has remedied the problem in its new standards.

In addition, the polygraph community has historically accepted any favorable research as evidence of validity. If the results looked good, the polygraph field was often eager to embrace it irrespective of uncomfortable questions about scientific methodology. This presented a not-to-be-missed opportunity for individuals to advance commercial or personal interests. One simply needed to issue a study that finds one's own techniques to be highly accurate, and the field accepted it as true.

Is this an exaggeration? If it were only so. There is ample evidence in the current polygraph literature of individuals doing exactly that, but the trend goes back nearly 100 years in the whole field of lie detection. The extraordinary accuracy reported in lie detection research conducted by parties with

a strong interest in the results is so striking, so unmistakable, it is perplexing why it has not been previously reported. Table 4 is a summary of the literature of reports where researchers reported on 1) the accuracy of their own methodologies, or 2) the accuracy of their decisions based on their own field cases. I was unable to locate a single study since 1914 meeting either of these two criteria that produced an accuracy more than 4% from perfection.

There are two possible interpretations for these findings of spectacular accuracy. One is that this body of research is valid, and that we can have confidence in the reports of perfect or near-perfect accuracy of the following methods: discontinuous blood pressure, the Pathometer, the Quadri-Track, the Arther, the CVSA, the Integrated ZCT, the Backster ZCT, brain waves, inhalation-exhalation ratios, and the Relevant/Irrelevant test.

Table 4. Summary of reported accuracy of individuals researching their own lie detection methods or using their own field cases, or both. Rounded to the nearest whole percent.

Researcher	Year	Technique	Accuracy
Marston	1921	Discontinuous Blood Pressure	**100%**
Summers	1936	Pathometer (EDA)	**100%**
MacNitt	1942	Relevant/Irrelevant	**100%**
Arellano	1984	Backster	**100%**
Matte	1989	Quadri-Track	**100%**
Benussi	1914	Pneumograph only	**100%**
Gordon, Mohamed, Faro, Platek, Ahmad & Williams	2005	Integrated ZCT	**100%**
Farwell	1993-2011	Brain waves	**100%**
Tippett	2004	CVSA	**100%**
Arther	1998	Arther	**100%**
Gordon, Fleisher, Morsie, Habib, & Salah	2000	Integrated ZCT	**100%**
Mangan, Armitage & Adams	2008	Quadri-Track	99%
Shurany & Chaves	2010	Integrated ZCT	99%
Putnam	1983	Backster & MGQT	99%
Edwards	1981	Various (Survey of 71 Examiners)	98%
Summers	1938	Pathometer (EDA)	98%
Shurany	2010	Quadri-Track	97%
Marston	1917	Discontinuous Blood Pressure	96%

A second possible interpretation is that practitioners need to be mindful of the potential conflict of interest when researchers self-evaluate their ideas and field data. This conflict, under the right conditions, can lead to exaggerated conclusions almost invariably in favor of the researcher. So, is there really such a thing as 100% accuracy in polygraph? Carl Sagan said it best: "Extraordinary claims require extraordinary evidence." The evidence for the virtually perfect polygraph technique does not rise to this standard.

Conclusion

The summary of twenty polygraph principles in this paper is aimed toward making polygraph examiners more conscious of factors that affect the validity of their chosen polygraph techniques. As new evidence is published, these principles will be further refined. The paper is also intended to help examiners choose from among the available techniques so that their practices will come to provide the most value to their departments, agencies or clients. It is also directed toward polygraph schools, so that they can bring their instruction in line with the current state of the evidence.

As the field of polygraphy approaches the completion of its first century it is showing distinct signs of maturity, such as the pursuit of best practices and an attentiveness to the scientific underpinnings of field methods. With that maturity has come an understanding of what the responsibility of a *profession* really is. It is not the single-minded protection of the industry, not solely the furtherance of economic interests, not a defensiveness against scrutiny: It is the protection of the public against the incompetent, the unethical, the poorly trained and the irresponsible practitioner, and the use of invalid methods. An important basis for fulfilling that professional duty is knowing which practices can be defended and which cannot. It is the author's hope that this summary can help examiners know the difference.

References

Aamodt, M. G. (2004). *Research in Law Enforcement Selection*. Brown-Walker Press: Boca Raton, FL.

Abrams, S. (1999). A response to Honts on the issue of the discussion of questions between charts. *Polygraph*, 29(3), 223-228.

Amsel, T. T. (1999). Exclusive or nonexclusive comparison questions: A comparative field study. *Polygraph*, 28(2), 273-283.

Anderson, N. B., & McNeilly, M. (1991). Age, gender, and ethnicity as variables in psychophysiological assessment: Sociodemographics in context. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 3(3), 376-384.

Arrellano, L. R. (1984). *Research paper: The polygraph examination of Spanish speaking subjects*. Unpublished manuscript.

Arther, R. O., & Arther, C. A. (1998). Truths about the proven truth verifier- the polygraph. *The Journal of Polygraph Science*, 33(3).

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

Backster, C. (1985). *Backster Zone Comparison Technique: Chart interpretation – Summary*. Paper presented at the 20th annual seminar of the American Polygraph Association, Reno, NV.

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

Benussi, V. (1914). Die atmungssymptome der lüge [The respiratory symptoms of lying]. *Archiv fuer die Gesamte Psychologie*, 31, 244-273. (Text in German). English translation published in *Polygraph*, 4 (1), 52-76.

Blackwell, N. J. (1999). PolyScore 3.3 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations. *Polygraph*, 28(2), 149-175.

Blalock, B., Cushman, B., & Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. *Polygraph*, 38(4), 281-288.

Bradley, M. T., & Cullen, M. C. (1993). Polygraph lie detection on real events in a laboratory setting. *Perceptual and Motor Skills*, 76, 1051-1058.

Buckley, J. P., & Senese, L. C., (1991). The influence of race and gender on blind polygraph chart analyses. *Polygraph*, 20(4), 247-258.

Crowe, M. J., Chimarys, M., & Schwartz, J. R. (1988). *The GQT polygraph test: Scoring and validity*. Poster presentation at the 96th annual convention of the American Psychological Association.

Cullen, M. C., & Bradley, M. T. (2004). Positions of truthfully answered controls on control questions tests with the polygraph. *Canadian Journal of Behavioural Science*, 36(3), 167-176.

Cushman, B., & Krapohl, D. J. (2010). *Evidence for technical questions in polygraph techniques*. Presentation at the APA Annual Seminar, Myrtle Beach, SC.

Dawson, M. E. (1981). Physiological detection of deception: Measurement of responses to questions and answers during countermeasure maneuvers. *Psychophysiology*, 17, 8-17.

Edwards, R. H. (1981). A survey: reliability of polygraph examinations conducted by Virginia polygraph examiners. *Polygraph*, 10(4), 229-272.

Farwell, L. A. (1993). *Brain-wave detection of concealed information*. Final report to the Office of Research and Development, Vol. 1. Human Brain Research Laboratory, Inc.: Potomac, MD.

Farwell, L. A. (2008). Brain fingerprinting detects real crimes in the field despite one-hundred-thousand-dollar reward for beating it. *Psychophysiology*, 45(s1), S1.

Farwell, L. A. (2011). Brain Fingerprinting: Corrections to Rosenfeld. *The Scientific Review of Mental Health Practice*, 8(2), 56-68.

Farwell, L. A. & Richardson, D. C. (2006a). Brain fingerprinting in laboratory conditions. *Psychophysiology*, 43(s1), S37–S38.

Farwell, L. A. & Richardson, D. C. (2006b). Brain fingerprinting in field conditions. *Psychophysiology*, 43(s1), S38.

Farwell, L. A. & Smith, S. S. (2001). Using brain MERMER testing to detect concealed knowledge despite efforts to conceal. *Journal of Forensic Sciences*, 46(1), 135–143.

Federal Psychological Detection of Deception Examiner Handbook (2006). Reprinted in *Polygraph*, 40(1), 1-66.

Franz, M. L. (1988). *Relative contributions of physiological recordings to detect deception*. Technical Report contract # MDA904-88-M-6612. Argenbright Polygraph, Inc.: Atlanta, GA.

Fuse, L. S. (1982). *Directed lie control testing technique*. Paper presented at the seminar of the Federal Interagency Polygraph Committee, FBI Academy, Quantico, VA.

Ginton, A., Daie, N., Elaad, E., & Ben-Shakhar, G. (1982). A method for evaluating the use of the polygraph in a real-life situation. *Journal of Applied Psychology*, 67(2), 131-137.

Gordon, N. J. (1999). The Academy for Scientific Investigative Training's Horizontal Scoring System and examiner's algorithm for chart interpretation. *Polygraph*, 28(1), 56-64.

Gordon, N. J., Fleisher, W. L., Morsie, H., Habib, W., & Salah, K. (2000). A field validity study of the integrated zone comparison technique. *Polygraph*, 29(3), 220-225.

Gordon, N. J., Mohamed, F. B., Faro, S. H., Platek S. M., Ahmad, H., & Williams, J. M. (2005). Integrated zone comparison polygraph technique accuracy with scoring algorithms. *Physiology & Behavior*, 87(2), 251-254.

Gustafson, L. A. & Orne, M. T. (1965). Effects of perceived role and role success on the detection of deception. *Journal of Applied Psychology*, 49(6), 412-417.

Handler, M., & Nelson, R. (2007). Polygraph terms for the 21st century. *Polygraph*, 36(3), 157-164.

Handler, M., Nelson, R., Goodson, W., & Hicks, M. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision polices. *Polygraph*, 39(4), 200-215.

Handler, M. D., Shaw, P. & Gougler, M. (2010). Some thoughts about feelings: A study of the role of cognition and emotion in polygraph testing. *Polygraph*, 39 (3), 139-154.

Harris, J. C., Horner, A., and McQuarrie, A. D. (2000). *An evaluation of the criteria taught by the Department of Defense Polygraph Institute for interpreting polygraph examinations*. SSD-POR-00-7272. Prepared under contract DABT02-96-C-0012-CLIN0004-Phase 11 for: The Department of Defense Polygraph Institute, Ft. Jackson, SC 29207, DoDPI00-R-0007. The John Hopkins University, Applied Physics Laboratory: Laurel, Maryland.

Hedges, K., & Deitchman, G. (2012). Does spot scoring and relevant and comparison question order help or hurt the examiner? A computer analysis of ground truth verified Army and Air Force MGQT and Federal ZCT exams. *Polygraph*, 41(3), 156-169.

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.

Honts, C. R. (1984). Countermeasures and the physiological detection of deception. *Psychophysiology*, 21, 566-567. (Abstract).

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

Honts, C. R. & Hodes, R. L. (1983). The detection of physical countermeasures. *Polygraph*, 12(1), 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(1), 177-187.

Honts, C. R., Raskin, D. C. & Kircher, J. C. (1983). Detection of deception: Effectiveness of physical countermeasures under high motivation conditions. *Psychophysiology*, 20, 446-447. (Abstract).

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

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

Horvath, F. S. (1988). The utility of control questions and the effects of two control question types in field polygraph techniques. *Journal of Police Science and Administration*, 16(3), 198-209. Reprinted in *Polygraph*, 20(1), 7-25.

Horvath, F. S., & Palmatier, J. 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), 889-899.

Khan, J., Nelson, R., & Handler, M., (2009). An exploration of emotion and cognition during polygraph testing, *Polygraph*, 38 (3), 184-197.

Kircher, J. C., Kristjansson, S. D., Gardner, M. K., & Webb, A. (2005). *Human and computer decision-making in the psychophysiological detection of deception*. Final report to the Department of Defense Polygraph Institute. University of Utah: Salt Lake City

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*. Report to the US. Department of Defense Polygraph Institute. DoDPI02-R-0002, DTIC AD Number A404128. Reprinted in *Polygraph*, 39(1), 22-66.

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

Kircher, J. C., Raskin, D. C., & Honts, C. R. (1984). Electrodermal habituation in the detection of deception. *Psychophysiology*, 21(5), 585. (Abstract).

Krapohl, D. J. (2005). Polygraph decision rules for evidentiary and paired-testing (Marin Protocol) applications. *Polygraph*, 34(3), 184-192.

Krapohl, D. J. (2010). Short report: A test of the ESS with two-question field cases. *Polygraph*, 39(2), 124-126.

Krapohl, D. J., & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: a replication. *Polygraph*, 35(1), 55-62.

Krapohl, D. J., & Dutton, D. W. (2005). A comparison of response profile for test formats used in the zone comparison and army modified general question techniques. *Polygraph*, 34(1), 1-9.

Krapohl, D. J., & Gary, W. B. (2004). Exploration into the effect of race on polygraph scores and decisions. *Polygraph*, 33(4), 234-239.

Krapohl, D. J, Gordon, N. J., & Lombardi, C. (2008). Accuracy demonstration of the Horizontal Scoring System using field cases conducted with the Federal Zone Comparison Technique. *Polygraph*, 37(4), 263-268.

Krapohl, D. J., McCloughan, J. B, & Senter, S. M. (2006). How to use the concealed information test. *Polygraph*, 35(3), 123-138.

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

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

Light, G. D. (1999). Numerical Evaluation of the Army Zone Comparison Test. *Polygraph*, 28(1), 37-45.

MacNitt, R. D. (1942). In defense of electrodermal response and cardiac amplitude as measures of deception. *Journal of Criminal Law, Criminology and Police Science*, 33(3), 266-275.

Mangan, D. J., Armitage, T. E., & Adams, G. C. (2008). A field study on the validity of the Quadri-Track Zone Comparison Technique. *Physiology & Behavior*, 95, 17–23.

Marston, W. M. (1917). Systolic blood pressure symptoms of deception. *Journal of Experimental Psychology*, 2(2), 117-163.

Marston, W. M. (1921). Psychological possibilities in deception tests. *Journal of the American Institute of Criminal Law and Criminology*, 11(4), 551-570.

Matte, J. A. (1999). Numerical scoring systems in the triad of Matte polygraph techniques. *Polygraph*, 28(1), 46-55.

Matte, J. A., & Backster, C. (2000). A critical analysis of Amsel 's comparative study of the exclusive v. nonexclusive comparison question. *Polygraph*, 29(3), 261-266.

Matte, J. A. & Reuss, R. M. (1989). A field validation study of the Quadri-Zone Comparison Technique. *Polygraph*, 18(4), 187-202.

Matte, J. A. & Reuss, R. M. (1990). A field study of the "friendly polygraphist" concept. *Polygraph*, 19(1), 1-8.

Matte, J. A., & Reuss, R. M. (1992). A study of the relative effectiveness of physiological data in field polygraph examinations. *Polygraph*, 21(1), 1-22.

Meijer, E., Verschuere, B., & Ben-Shakhar, G. (2011). Practical guidelines for developing a CIT. In Verschuere, Ben-Shakhar & Meijer (Eds.) *Memory detection: Theory and application of the Concealed Information Test.* Cambridge University Press: New York.

Menges, P. M. (2004). Directed lie comparison questions in polygraph examinations: History and methodology. *Polygraph*, 33(3), 131-142.

Miyake, Y. (1978). A study of skin resistance response, photoplethysmograph vasomotor response and eye movement as indices of lie detection. *Reports of the National Research Institute of Police Science*, 31(2), 18-24.

National Research Council (2003). National Research Council (2003). *The Polygraph and Lie Detection. Committee to Review the Scientific Evidence on the Polygraph*. Washington, DC: The National Academies Press.

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B, & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph*, 40(2), 67-78.

Nelson, R., Krapohl, D. J., & 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(3), 185-215.

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(3), 145-155.

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

Orne, M. T. (1973). Implications of laboratory research for the detection of deception. *Polygraph*, 2(3), 169-199.

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

Peters, R. (2011). A history of polygraph digitization: Credibility sleuths encounter the geeks. *Polygraph*, 40(3), 166-171.

Podlesny, J. A. (1993). Is the guilty knowledge polygraph technique applicable in criminal investigations?: A review of FBI case records. *Crime Laboratory Digest*, 20(3), 57-61.

Podlesny, J. A., & Kircher, J. C. (1999). The Finapres (volume clamp) recording method in psychophysiological detection of deception examinations. *Forensic Science Communications*, 3(1). Available at: http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/oct1999/podlsny1.htm.

Podlesny, J. A., Nimmich, K. W., & Budowle, B. (1995). *A lack of operable case facts restricts applicability of the guilty knowledge deception detection method in FBI criminal investigations: A technical report.* Federal Bureau of Investigation Forensic Science Research and Training Center. Quantico, Virginia.

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

Putnam, R. L. (1983). Field accuracy of polygraph in the law enforcement environment. *Polygraph*, 23(3), 260. (Abstract).

Raskin, D. C. (1976). *Reliability of chart interpretation and sources of error in polygraph examinations*. Report No. 76-3, National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice (Contract No. 75-NI-99-0001). Department of Psychology, University of Utah.

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., 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. University of Utah: Salt Lake City, UT.

Reed, S. (1993a). *Effect of demographic variables on psychophysiological detection of deception examination outcome accuracies*. DoDPI93-R-0007. Department of Defense Polygraph Institute, Ft. McClellan, AL. DTIC AD Number A304664.

Reed, S. (1993b). *Subcultural report -- Effects of examiner's and examinee's race on psychophysiological detection of deception outcome accuracy*. DoDPI94-R-0012. Department of Defense Polygraph Institute, Ft. McClellan, AL. DTIC AD Number A310901.

Reid, J. E. (1945). Simulated blood pressure responses in lie detector tests and a method for their detection. *American Journal of Police Science*, 36(1), 201-214.

Research Division Staff (1995a). *A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope polygraph and the test for espionage and sabotage question formats*. Department of Defense Polygraph Institute, DoDPI94-R-0008. Ft. McClellan, AL.

Research Division Staff (1995b). *Psychophysiological detection of deception accuracy rates obtained using the test for espionage and sabotage*. Department of Defense Polygraph Institute, DoDPI94-R-0009. Ft. McClellan, AL.

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

Senter, S. M. (2003). Modified general question test decision rule exploration. *Polygraph*, 32(4), 251-263.

Senter, S. M., & Dollins, A. B. (2008). Comparison of question series and decision rules: A replication. *Polygraph*, 33(4), 223-233.

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. M., Dollins, A. B., & Krapohl, D. J. (2004). A comparison of polygraph data evaluation conventions used at the University of Utah and the Department of Defense Polygraph Institute. *Polygraph*, 33(4), 214-222.

Senter, S. M., Weatherman, D., Krapohl, D. J., & Horvath, F. (2010). Psychological set or differential salience: A proposal for reconciling theory and terminology in polygraph testing. *Polygraph*, 39(2), 109-117.

Shurany, T., Stein, E., & Brand, E. (2009). A field study on the validity of the Quadri-Track Zone Comparison Technique. *European Polygraph*, 1(7), 5-23.

Stern, R. M., & Kircher, J. C. (2002). *The effects of augmented physiological feedback on detection of deception*. Report to the DoD Polygraph Institute. DoDPI02-R-0009. Reprinted in *Polygraph*, 39(4), 216-232.

Stewart, W. S. (1941). How to beat the lie detector. *Esquire*, 16(5), 35, 158, 160.

Summers, W. G. (1936). A recording psychogalvanometer. *Bulletin of American Association of Jesuit Scientists, Eastern States Division*, 14(2), 50-56. Reprinted in *Polygraph*, 13(4), 340-345.

Tippett, R. G. (2004). *Comparative analysis study of the CVSA and polygraph*. Unpublished (but widely cited) manuscript.