Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

VOLUME 47 **2018** **NUMBER 2**

Contents

Published Semi-annually

P.O. Box 8037, Chattanooga, Tennessee 37414-0037

Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

Editor-in-Chief: Mark Handler
E-mail: Editor@polygraph.org
Managing Editor: Nayeli Hernandez
E-mail: polygraph.managing.editor@gmail.com

******

Subscription information: *Polygraph* is published semi-annually by the American Polygraph Association. Editorial Address is Editor@polygraph.org. Subscription rates for 2018: One year $150.00 (Domestic), $180.00 (Foreign). Change of address: APA National Office, P.O. Box 8037 Chattanooga, TN 37414-0037. THE PUBLICATION OF AN ARTICLE IN *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* DOES NOT CONSTITUTE AN OFFICIAL ENDORSEMENT BY THE AMERICAN POLYGRAPH ASSOCIATION.

Instructions to Authors

Scope

The journal *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* publishes articles about the psychophysiological detection of deception, and related areas. Authors are invited to submit manuscripts of original research, literature reviews, legal briefs, theoretical papers, instructional pieces, case histories, book reviews, short reports, and similar works. Special topics will be considered on an individual basis. A minimum standard for acceptance is that the paper be of general interest to practitioners, instructors and researchers of polygraphy. From time to time there will be a call for papers on specific topics.

Manuscript Submission

Manuscripts must be in English, and may be submitted, along with a cover letter, on electronic media (MS Word). The cover letter should include a telephone number, and e-mail address. All manuscripts will be subject to a formal peer-review. Authors may submit their manuscripts as an e-mail attachment with the cover letter included in the body of the e-mail to:

Editor@polygraph.org

As a condition of publication, authors agree that all text, figures, or other content in the submitted manuscript is correctly cited, and that the work, all or in part, is not under consideration for publication elsewhere. Authors also agree to give reasonable access to their data to APA members upon written request.

Manuscript Organization and Style

All manuscripts must be complete, balanced, and accurate. Authors should follow guidelines in the *Publications Manual of the American Psychological Association*. The manual can be found in most public and university libraries, or it can be ordered from: American Psychological Association Publications, 1200 17th Street, N.W., Washington, DC 20036, USA. Writers may exercise some freedom of style, but they will be held to a standard of clarity, organization, and accuracy. Authors are responsible for assuring their work includes correct citations. Consistent with the ethical standards of the discipline, the American Polygraph Association considers quotation of another's work without proper citation a grievous offense. The standard for nomenclature shall be the *Terminology Reference for the Science of Psychophysiological Detection of Deception* (2012) which is available from the national office of the American Polygraph Association. Legal case citations should follow the *West* system.

Manuscript Review

An Associate Editor will handle papers, and the author may, at the discretion of the Associate Editor, communicate directly with him or her. For all submissions, every effort will be made to provide the author a review within 4 weeks of receipt of manuscript. Articles submitted for publication are evaluated according to several criteria including significance of the contribution to the polygraph field, clarity, accuracy, and consistency.

Copyright

Authors submitting a paper to the American Polygraph Association (APA) do so with the understanding that the copyright for the paper will be assigned to the American Polygraph Association if the paper is accepted for publication. The APA, however, will not put any limitation on the personal freedom of the author(s) to use material contained in the paper in other works, and request for republication will be granted if the senior author approves.

Believing Is Seeing: The Influence of Expectations on Blind Scoring of Polygraph Data

Donald J. Krapohl[1] and Donnie W. Dutton[2]

Abstract

Modern polygraph scoring systems were developed to give structure to polygraph chart interpretation with the aims of increasing accuracy and inter-examiner agreement while decreasing outside influences on chart interpretation. Previous research shows high agreement among polygraph examiners' decisions when conducting blind scoring of polygraph data (Blackwell, 1999). However, the evidence does not support the presumption that manual scoring systems make scorers resistant to bias. In previous work by Elaad, Ginton and Ben-Shakhar (1994) and Shurany, Matte and Stein (2009) researchers found that examiners' scores and decisions could be influenced by what researchers told the examiners about the cases before the examiners scored them. To further explore the effect of outside influence on polygraph scoring we recruited polygraph examiners through a large polygraph professional organization to blindly evaluate 15 polygraph cases, some of which were ambiguous and others which clearly indicated truthfulness or deception. Half of the polygraph examiners were told that all of the cases were confirmed as deceptive while the other half were told all of the cases were verified as truthful. We found, on average, polygraph scores and decisions were shifted in the direction of the biasing information. The shift was evident for both clear and ambiguous data. Not all scorers were affected by the biasing information. Implications are discussed.

Keywords: 7-position scoring, agreement, bias, decision making, expectation, reliability

All humans come to the task of decision-making with biases. Biases *per se* are neither bad nor good. As an example of contrast, having a bias toward one's own offspring is probably an important part of the survival of a species or a family line: Having a bias toward unhealthy eating works against this same goal. The badness and goodness of biases then are generally judged by their effects and by which parties are affected.

Most professions are aware that practitioner biases can lead to sub-optimal outcomes. One method for controlling bias is the use of evaluation methodologies that rely more on objective measures such as scoring systems and algorithms. For the polygraph, scoring arrived many years after the field was established. In its first decades, evaluation of polygraph data was almost entirely subjective. There was no common method of analysis among practitioners, though some general ideas for data interpretation were available by the early 1930s (Larson, 1932). The first published quantification system for polygraph "lie detection" data was suggested by Winter

1 APA Past President and regular contributor to this publication. Comments regarding this project should be sent to him at APAkrapohl@gmail.com.

2 APA Past President and regular contributor to this publication.

We are grateful for the assistance of APA Editor-in-Chief Mark Handler for facilitating the release of email announcements used to recruit volunteer scorers. His help was critical to the success of the study.

(1936). Winter's method was simple and entailed tallying Xs he had made in a table to indicate when tests showed indications of guilt. The Winter method appeared in only one publication and was not adopted by field practitioners. A somewhat related approach called the "checkmark system" was formally used by polygraph pioneers John Reid, Richard Arther and their students before 1960 (Slowik & Horvath, 2016). However, their decision-making also incorporated non-polygraph information such as case facts and impressions of examinee behavior, clinical observations that can resist purely objective quantification.

Numerical scoring of polygraph charts was an innovation introduced to the polygraph profession by Cleve Backster in 1959 (Matte, 1996, citing Backster, 1969). Backster's 7-position scoring system was an adaptation of the Likert Scale (Likert, 1932) though the historical record is silent on whether Mr. Backster independently developed his method. At its most basic level the Backster scoring system was a straightforward accounting method that tracked the relative frequency and intensity of differences between physiological reactions to relevant questions[3] and comparison questions[4] under the rationale of the comparison question polygraph test (Kircher & Raskin, 1988; Kubis, 1962; Reid, 1947; Summers, 1939) [see Nelson (2016) for a discussion]. Backster's 7-position scoring system was adopted by the US government in the early 1960s and quickly spread across the polygraph profession until Backster's method, or variations of it, became the dominant approach to polygraph chart evaluation. Numerical scoring is considered one of the most significant advances in polygraph examinations.

Raskin, Barland and Podlesny (1977) confirmed that the use of 7-position scoring improved decision accuracy over the use of impressionistic methods. The reduction in subjectivity inherent to a shared method of analysis does not automatically permit a conclusion that all sources of polygraph bias are removed. Biological signals are complex, rich and notoriously variable. As such they provide fertile opportunities for subjective individual differences in data evaluation. Biases can be unwittingly introduced by factors such as prior experience, culture, training, halo effect, or expectations.

Among these potential sources of bias in polygraph chart evaluation only examiner expectations have undergone any published investigation. Because they can be manipulated with relative ease expectations lend themselves well to bias studies. Elaad, Ginton and Ben-Shakhar (1994) reported two experiments wherein they accurately or inaccurately informed polygraph examiners of ground truth before asking them to conduct blind evaluations of field polygraph charts. In both experiments Elaad et al. used 10 polygraph examiners employed by the Israel National Police (INP) who were experienced with numerical scoring to independently evaluate 14 sets of criminal field polygraph charts. In the first experiment Elaad selected polygraph cases that had been blind scored previously by three independent examiners and judged unanimously or by majority to be inconclusive. In other words, the physiological data would not clearly lead to a definitive result of truthfulness or deception. The 14 cases were then given to INP examiners with instructions to score the charts. To manipulate expectations, examiners were told that either the cases came from examinees who confessed to the crime for which they were being polygraphed or that the charts were from cases where the examinee was exculpated by confession of someone other than the examin-

3 Relevant questions cover the issue(s) of interest to the client or agency. They are the reason for the examination.

4 Comparison questions are those designed to elicit physiological reactions, and generally regard whether the examinee has engaged in anti-social or proscribed behaviors in the past, though there are variations to the question content. In polygraph scoring the reactions to comparison questions are gauged against the reactions of relevant questions at 18 to 45 or more places in the polygraph charts. Greater reactions to comparison questions warrant a positive score at each of the scoring locations and greater reactions to relevant questions are assigned a negative score. The scores are tallied at the end of the examination and the totals are compared to decision thresholds to form a conclusion of truthfulness or deception.

ee. When the examiners' scores were evaluated by the experimenters they found that they had shifted significantly in the direction of the expectation. That is, when examiners believed they were scoring the charts of deceptive examinees their scores shifted significantly in the negative direction, and when the expectations were of truthfulness the examiners assigned scores that shifted positively. The level of examiner experience was not a factor in the shift.

In their second experiment Elaad used the exact same procedures and examiners. However, in this experiment they selected a different set of polygraph cases in which the data allowed definitive results of deception or truthfulness rather than the ambiguous cases used in the first experiment. While the examiners' scores shifted again in the predicted direction for the second experiment, the effect was not statistically significant. The researchers concluded that when polygraph data are unambiguous polygraph scores are not significantly affected by expectations, whereas the scores shift in the direction of expectations when the polygraph data are not clear.

In 2009, Shurany, Matte and Stein used a variation of the Elaad manipulation again to assess the influence of expectations on examiner decisions. During training events they used 82 experienced polygraph examiners from three countries; Mexico, United States and Bulgaria. They projected polygraph charts from eight cases individually onto a screen and asked the examiners in attendance to individually score them using the 3-position scoring system, an abbreviated form of 7-position scoring system devised by Backster. The first and eighth cases were actually the same case, one that had been confirmed as truthful, though the examiners were not informed as such. When the case was presented the first time the instructor provided no information about the case background. Fictitious case facts were provided to the scorers just before scoring the second through the seventh cases. When the first case was presented again as the eighth case the instructor falsely reported that a fingerprint was found at the scene of the crime that had a 90% match with the examinee's print. Shurany et al. then compared the examiners' scores for the repeated case to determine the effect of the fingerprint information. They found that when examiners evaluated the case after learning of the fingerprint their averaged scores moved significantly in the direction of deception though the case in reality was from a truthful examinee. Decision accuracy was also affected. In the initial showing of the case, with no background information, only one scorer out of 82 had scored the case as deceptive. This increased to nine scorers after the false information was implanted, a statistically significant difference. Shurany also discovered that the biasing effect appeared among scorers from all three represented countries. In the report's conclusion Shurany recommended independent and blind quality control to mitigate the potential of bias that could be experienced by the testing polygraph examiner.

A significant advantage of the Shurany study was the use of a large and heterogenous sample which might have served to increase generalization of his findings. Shurany showed a significant biasing effect on polygraph scores and decisions that manifest itself among examiners from three separate countries. The study's chief limitation was the use of a single case to base its conclusion of bias. The effect on the single truthful case did not address questions of repeatability, nor whether the effect would be found with deceptive cases. Advantages of the Elaad studies were that they used 14 cases in each of their manipulations and varied the data between ambiguous and clear. However, the Elaad sample was quite homogenous, where all 10 polygraph examiners worked for one police organization in Israel. Whether their conclusions would also be found among examinees elsewhere is not known.

It became our interest to assess the influence of expectations on as broad a range of field examiners as reasonably possible using a sample of polygraph cases that would permit the investigation of differences in bias brought about by clear or by unclear polygraph data.

Method

Ethical Considerations

With regard to securing consent from the scorers in the study, an explanation of the scoring task was sent via email to all practicing polygraph examiners with the American Polygraph Association. Participation required that they respond to the email expressing a willingness to score field charts using a specified method within a finite amount of time. Those who responded positively to that email were sent charts for scoring. The subset of examiners who expressed a willingness to participate and who ultimately submitted their data became our sample. Consent was implied by these voluntary steps undertaken by the scorers.

Because the present study entailed the manipulation of expectations we recognized that deception of participants would be necessary. We referred to the ethical standards of the American Psychological Association (2017), Section 8.07, that deception in human research should be avoided except when: 1. Research cannot be conducted without deception; 2. The research may help answer important questions, and; 3. Those to be deceived should not be harmed by the deception.

Regarding Standard 1, whether deception was necessary, we know of no practical means of creating different expectations between groups without deceiving one or more of the groups. We concluded that deception was unavoidable for this research question. Two different deceptions would be necessary to ensure the data would be meaningful, one of the true purpose of the study and the other the manipulation of expectations.

For Standard 2, and whether the research question is sufficiently important, we accept that there may be debate as to the worthiness of any research undertaking. An important factor in our decision to proceed was the impact polygraph examinations usually have on the lives of those who undergo the testing. The results of polygraph testing can be impactful on decisions for hiring, prosecutions, probation revocation, security clearances and criminal investigations. Whether polygraph decisions are biased by examiner expectations has real world import sufficient to justify its investigation in our view.

To the final consideration regarding harm[5] to those who are deceived, we took exceptional steps to minimize the potential for harm. Though no physical injury could be expected from our study, we were mindful of the potential for creating stress for examiners when they discovered our manipulation of expectations. We judged this to pose a minimal risk for this group. First, we concluded from our extensive contact with polygraph professionals, as well as being longtime practitioners ourselves, that experienced polygraph examiners manage stress well, are accustomed to disagreement, and that their self-esteem trends toward robust. We concluded that, in the main, polygraph examiners are not an at-risk population for modest stress. We also recognized that polygraph examiners undergo assessments of their ability to score charts from the initial weeks of their polygraph education through feedback during their professional practices. Not all of that feedback is favorable, and some of it can be unsettling. We considered that experienced polygraph examiners receiving positive or adverse feedback regarding their performance with the study charts would not view it more stressful than their daily interpretation of polygraph charts that had significant impact on the examinee. Second, we anonymized all of the data so that no scores could be attributed to any given scorer. This was done in the initial processing of the data, and there is no subsequent means for matching scorer with scores. All subsequent reporting was of group data. Finally, following data collection we sent a debriefing email to all participants explaining the purpose of the study, the need for misdirection, the importance of the project, that all scores had been anonymized, that all reporting would be of group data alone, and we provided copies of the two previous bias studies discussed earlier to place our study in context. Contact information to the primary researcher was included in that communication. Following our email notification a small number of scorers

5 The language used for this ethical standard regarding harm to participants was that the research would be ..."reasonably expected to cause physical pain or severe emotional distress."

responded, but none expressed displeasure with the manipulation nor requested withdrawal of their data.

Participants

A total of 73 scorers submitted scores for the study. They self-reported that they worked in one of three sectors: private practice, law enforcement or government. Examiners in private practice represented the largest group of volunteers with 44 (60.3%) submitting data, followed by law enforcement with 25 (34.2%) and then government with 4 (5.5%). Examiners within the sector categories were not as homogenous as one might suspect, however. Among polygraph examiners in private practice a portion were retired from government or law enforcement while others had no experience in the public sector. Some private sector examiners also did contract polygraph testing for law enforcement or government. The four self-described government examiners worked for three different governments. Consequently, there may be few examiner characteristics within a category that are unique to that category or common within the category.

There was a very wide range of experience, from 6 months to 50 years. The mean number of years was 18.5 (sd = 13.8). Other demographics such as examiner age, gender and ethnic group were not solicited or captured.

Polygraph Cases

Fifteen cases were selected from the first author's archive of field polygraph examinations accumulated over 12 months of performing quality control reviews of sex offender examinations for a large program. All 15 cases were conducted using the Air Force Modified General Question Technique with two relevant questions, variation 1 (Krapohl & Shaw, 2015) and had three charts. None of the cases were confirmed as truthful or deceptive. Five of the cases were chosen because the original examiner and the quality control reviewer found strong indications of truthfulness in the polygraph data. The mean score for these cases by the testing examiner was +12.8 (sd = 3.35) and +11.4 (sd = 3.05) for the quality control reviewer. A second group of five were selected because the testing examiner and quality control reviewer agreed the charts strongly indicated deception. The mean score for the testing examiner for these cases was -10.6 (sd = 5.72) and -10.0 (sd = 4.85) for the quality control reviewer. A final group of five cases were inconclusive according to the testing examiner and quality control reviewer. Average score for this last group of cases was +1.0 (sd = 5.15) for the testing examiner and -0.4 (sd = 5.50) for the quality control reviewer. All polygraph charts had been recorded on Lafayette computer polygraphs.

The charts were converted to PDF. The choice to use PDF rather than sending the electronic Lafayette files to scorers was based on three considerations. First, it was important to standardize the recordings as much as possible. Fixing the charts in PDF restricted the scorers' ability to change any of the gain settings and therefore all of the scorers would be evaluating the exact same data. Second, it prevented scorers from employing any of the algorithms available in the Lafayette software that might have influenced their scoring. Finally, though Lafayette computer polygraphs are used by many examiners, it was not expected that all scorers would have access to that manufacturer's systems. In contrast, software for viewing PDFs is universally available and at no cost to the user. It also allowed scorers to select and magnify portions of the charts to the degree necessary to assign scores. All charts contained standard question labels so examiners could determine which questions to score. The test questions themselves were not included in the files sent to scorers.

The 15 sets of charts were placed into a single PDF file, one case per page. All examiners received the same cases, but the order of cases was different between the truthful-expectation and deception-expectation groups. This was done to minimally hide the fact that both groups had received the same cases. The first page of the PDF of cases included instructions. Half of the volunteers received an instruction page indicating that all 15 cases were confirmed as deceptive, while the remainder of the volunteers were informed the cases came from verified truthful examinees. The instruction pages were identical for both groups except the words "deceptive" or "truthful" where ground truth was discussed. The instruction page for the truthful expectancy is found in Appendix A. Examiners were also provided with an Excel spreadsheet to record their scores. The spreadsheet likewise included information consistent with the expectation manipulation.

Polygraph Chart Scoring

There are several named manual scoring systems to analyze polygraph charts. To allow direct comparison among scorers we required they all be familiar with and use the 7-position scoring system as taught by the National Center for Credibility Assessment (NCCA). NCCA's scoring systems have been taught at most polygraph schools for decades. All volunteers were sent detailed instructions on NCCA's approach to 7-position scoring to help standardize the methods of all volunteer scorers.

Procedure

With the assistance of the American Polygraph Association a broadcast email was sent to its practicing members to seek volunteers willing to participate in a blind scoring project. The pretended purpose of the study was to determine which of the polygraph's four data channels provided the most diagnostic information. A total of 102 examiners responded electronically to the initial invitation. They were provided details of the study in a follow-on email (Appendix B) with a request for information regarding the sector in which the examiners worked, their years of polygraph experience, and the state (or country) where they practiced.

The volunteer scorers were assigned into deception-expectation and truthfulness-expectation groups randomly, with one modification. Examiners working in the same state or foreign country were programmed with the same expectancy, either deception or truthfulness. This procedure was put into place to help mitigate the risk that geographically close examiners would discover they were scoring the same cases but had received opposite information on "ground truth."

The 102 volunteers were sent electronic copies of polygraph charts to score. The scorers were sent only the PDF files of charts, participant instructions and an Excel spreadsheet file to enter scores. They were not given background information, the test topics or the test questions. From the 102 examiners who expressed initial interest in the study a total of 73 (71.5%) submitted scoring data. The truthful-expectation group produced 39 sets of scores and the deception-expectation group submitted 34 sets. A test of proportions found no significant difference in the return rate of scores between these groups (z = 0.70, ns).

All data were anonymized using numbers in place of names of the scorers. Only the scorers' years of experience and sector of practice were included with the scores.

We used cutoff scores of +/-4 total for the decision threshold in the study. This was because the original examiner and quality control reviewer had had a minimum of +/-4 for the 10 cases in which they made definitive decisions. These cutoff scores were arbitrary and do not match those used by most examiners in the field.

Results

Decisions

The truthful-expectation (TE) and deception-expectation (DE) groups were compared to one another for the respective number of decisions of Deception Indicated (DI), No Deception Indicated (NDI) and Inconclusive (INC) using the imposed cutoff scores of +/-4 per case. With 15 possible decisions for each of the two groups, the TE group average 7.4 decisions (sd = 2.72) of NDI compared to 3.2 (sd = 2.14) decisions of NDI for the DE group. The difference was significant ($t[71] = 7.06$, $p = .000$). The TE group averaged 4.0 decisions (sd = 1.78) of DI overall compared to 7.8 (sd = 2.74) decisions of DI for the DE group. This difference was also significant ($t[71] = 6.99$, $p = .000$). There were no significant differences between the number of INC decisions from the TE and DE groups ($t[71] = 0.85$, ns). Table 1 shows the average number of DI, NDI and INC decisions for the TE and DE groups.

**Table 1. Average number of NDI, DI and INC decisions for the truthful-expectation and deception-expectation groups for the 15 cases they reviewed. *** $p < 0.05$.

	Truthful Expectation	Deception Expectation	Difference
NDI Decisions	7.4	3.2	4.2*
DI Decisions	4.0	7.8	3.8*
INC Decisions	3.6	4.0	0.4
Total	15	15	

Scorer expectations also influenced their agreement with the decisions of the original polygraph examiner. As Table 2 shows, groups with different expectations of ground truth tended to adjust their decisions to be in accord with the expectation. The TE group, which had been informed that all 15 of the cases had come from confirmed truthful examinees, had high concordance (91.2%) with the subset of five original decisions of truthfulness by the testing examiner. The DE group, expecting that all cases were confirmed as deceptive, were significantly less likely to agree (50.6%) with the original examiner on the same cases. Conversely, the DE group had 91.2% agreement with the decision of the original examiner when the examiner had decisions of deception on those five cases, but the TE group only agreed in about two-thirds of the time. There was a similar effect on the five cases the testing examiner had called Inconclusive, with TE and DE groups significantly diverging in their decisions.

Table 2. Rates at which groups with truthful expectations and deception expectations (rows) made decisions of NDI, DI and Inconclusive (INC) for cases independently deemed as truthful, deceptive or inconclusive (columns). Values are in percentage. All differences between groups significant at $\alpha < 0.05$ **are denoted with an *.**

	Truthful Cases			Deceptive Cases			Inconclusive Cases		
Decisions	NDI	DI	INC	NDI	DI	INC	NDI	DI	INC
Truthful Expectation	91.2	1.0	7.6	12.8	66.6	20.6	42.6	12.8	44.6
Deception Expectation	50.6	17.0	32.4	1.8	91.2	7.0	11.8	47.0	41.2
Absolute Difference	40.6*	16.0*	24.8*	11.0	24.6*	13.6	30.8*	34.2*	3.4

Table 2 suggests a large proportion of scorer disagreement were represented by the scorer making decisions of Inconclusive when their expectations did not match what they were seeing in the polygraph data. What may not be immediately obvious in the table is that scorer decisions opposite the original decisions represent a minority. To investigate how often this occurred we conducted an analysis of the data. Table 3 summarizes how often scorers were making these opposite decisions. Among those expecting truthfulness, 23 out of 39 (59%) of the scorers made no opposite decisions and 16 out of 34 (47%) of those expecting deception had no opposite decisions from the original examiner. Therefore, it might be concluded that on average about half of the scorers were not persuaded by the manipulation to make any decisions opposite from what the polygraph data indicated. Rather, they tended to place those cases into the Inconclusive bin in higher proportions. Less comforting, however, the other half of the scorers showed a willingness in varying degrees to form decisions in the direction of their expectations and entirely contrary to the polygraph chart data.

Table 3. Number of scorers who made 0, 1, 2, 3, 4 or 5 decisions directly opposite those of the original testing examiner. No scorer made more than 5 opposite decisions. None of the proportions in each of the six categories (columns) in the table were significantly different between expectation groups (rows).

		Number of Scorers Who Made Decisions Opposite to the Original Results					
		0	1	2	3	4	5
Expectation	Truthful	23	10	3	1	2	0
	Deception	16	10	4	3	0	1

Of some practical relevance to this project is what a commonly available polygraph algorithm would have decided with these same cases. The Objective Scoring System version 3 (OSS-3; Nelson, Krapohl & Handler, 2008) analyzed the 15 cases used in this project. OSS-3 was not accessed by the original examiner to formulate any decisions and it was not used as part of the selection process for any of the 15 cases in this study. The algorithm was in perfect agreement with the original examiner on decisions of truthfulness and deception but was more aggressive in its decisions for the 5 cases called Inconclusive by the original examiner. Of those five Inconclusive cases the algorithm made four definitive decisions. Because ground truth is not known for any of the cases in this project, accuracy could not be calculated. Table 4 shows the distribution of decisions, which are represented as percentages to permit direct comparison to Table 2.

Table 4. Rates at which the OSS-3 algorithm made decisions of NDI, DI and Inconclusive (INC) for cases independently deem as truthful, deceptive or inconclusive. Values in percent.

	Truthful Cases			Deceptive Cases			Inconclusive Cases		
Decisions	NDI	DI	INC	NDI	DI	INC	NDI	DI	INC
OSS-3 algorithm	100	0	0	0	100	0	20	60	20

Scores

As might be anticipated from the previous Decisions section, polygraph scores also shifted according to scorer expectations. Table 5 lists the average mean scores for truthful, deceptive and inconclusive cases by expectation. For the five truthful cases the TE group averaged 10.41 (sd = 2.90) points per case whereas the DE group averaged only 3.68 (sd = 4.53) per case. This difference is significant (t[71] = 6.74, p = 0.000). A similar shift in scores can be observed for the five deceptive cases. The TE group average -6.25 (sd = 4.73) points per case compared to -13.42 (sd = 4.29) points by the DE group, also significantly different (t[71] = 7.66, p = 0.000). For the inconclusive cases the TE group also assigned more positive scores per case at 2.23 (sd = 3.33) against a -3.33 (sd = 3.05) for the DE group. This difference was also significant (t[71] = 7.40, p = 0.000).

Table 5. Average total scores per case for TE and DE groups for cases previously deemed truthful, deceptive or inconclusive.

	Truthful Cases	Deceptive Cases	Inconclusive Cases
Truthful Expectation	10.41	-6.25	2.23
Deception Expectation	3.68	-13.42	-3.33
Absolute Difference	6.73	7.17	5.56

Bias Index

To assess the degree of bias we created a simple "Bias Index" which gauged scorers' decisions against the results of the original examiner. In the Bias Index a scorer began with 0 points. If the scorer made a decision in the direction of the expectation manipulation and opposed to the original examiner decision the scorer received 2 points. For example, if a scorer had an expectation of deception and called 3 of the 5 "truthful" cases as deceptive, that scorer would receive 6 bias points (2 points X 3 decisions opposite those of the original examiner.) With a total of five cases in each category of NDI and DI a scorer in either expectation group could garner a maximum of 10 points for decisions that were diametrically opposite those of the original examiner. If instead the scorers had made Inconclusive decisions on any of these same five cases, they would receive 1 point for each of the cases instead of 2 points. In other words, an opposite result was worth 2 points but a shift only to an inconclusive result added only 1 point each to the Bias Index. Finally, if the scorer made decisions on the 5 original "Inconclusive" cases in the direction of the expectation manipulation, the scorer would receive 1 point per case. A scorer could receive up to 15 points in this Bias Index. A final score of 0 points in the index would suggest the scorer was not influenced by expectations and a final score of 15 points would indicate perfect influence of expectations.

The average bias score for the TE group was 4.44 points (sd = 3.49) and a range of 0 to 13 points. The average bias score for the DE group was 5.58 points (sd = 3.72) with a range of 0 to 15 points. A t-test was conducted to determine whether the bias scores of the two groups were different from one another, which found they were not (t(71) = 0.15, ns). There were also no significant correlations between years of polygraph experience and bias score for either the TE group (r = 0.001, ns) or the TD group (r = 0.13, ns).

For convenience of discussion later in this paper we divided the bias scores into three nominal categories: 0 – 5 points = low bias, 6 – 10 points = moderate bias, and 11 – 15 = high bias. Table 6 lists the number of scorers who fall into these categories. For both TE and DE groups most of the scorers fell into the low bias category. Tests of proportions found no significant differences between the expectation groups for the proportions of scorers in each of the bias three categories.

Table 6. Number of scorers (and percentages) in each of the bias categories for the truthful and deception expectation groups.

	Truthful Expectation	Deception Expectation
Low bias (0 - 5 bias points)	27 (69.2%)	18 (52.9%)
Moderate bias (6 - 10 bias points)	8 (20.5%)	12 (35.3%)
High bias (11 - 15 bias points)	4 (10.3%)	4 (11.8%)
Total	39 (100.0%)	34 (100.0%)

Discussion

Our findings partially replicated those of Elaad et al. (1994) in their assessment of polygraph examiners with the Israel National Police. Consistent with Elaad, cases originally considered inconclusive by the testing examiner and the quality control reviewer tended to be called deceptive or truthful by our blind scorers in the direction of the expectation manipulation. The two studies taken together support the conclusion that when the polygraph data are unclear scorers appeared to be affected by expectations. The *a priori* belief in the examinees' veracity seemed to have led the two expectancy groups to evaluate the same data in sharply divergent ways.

We did not anticipate the same tendency among scorers when the polygraph charts clearly indicated truthfulness or deception, though the Shurany et al. (2009) single-sample data hinted at such a finding. Our data showed the original testing examiner, the independent quality control reviewer and subsequently a polygraph algorithm all agreed that 5 of the 15 cases plainly indicated truthfulness and a second group of 5 were definitely deceptive cases. As seen in Table 2, the scorers expecting deceptiveness agreed at a very high rate on the five cases originally considered deceptive, but only half as often on cases originally called truthful. Similarly, the scorers receiving the truthful manipulation did exceptionally well with the cases originally called truthful, but only for about two-thirds of the cases originally called deceptive.

Implications

Taken in to, it would appear our simple manipulation of expectations had a significant effect on the interpretation of not only ambiguous polygraph data, but on clear-cut cases, as well. On its face this could be unwelcomed news to those who rely on polygraph services. It is important to note, though, there may be important differences between the conditions of the study and conditions in the field that could affect generalization of the findings. Among these differences are that many polygraph examiners in private practice, most in law enforcement, and all government polygraph examiners have independent quality control built into their decision processes. It typically includes blind scoring of the polygraph data, an important step to diminish the effect of bias. An independent review of the polygraph data can provide not only the detection of, but deterrence against, bias in the scoring of polygraph data.

In addition, all modern polygraphs are computerized and include algorithms for the analysis of the data. These automated analytical systems are immune to social influences or extra-polygraphic information that may affect a scorer and are routinely used by many field examiners. Indeed, every study that has compared human decisions with algorithm decisions has found a small but consistently better performance for the algorithm decisions, regardless of which algorithms were employed (Blackwell, 1999; Kircher & Raskin, 1988; Kircher, Kristjansson, Gardner & Webb, 2005; Nelson, Krapohl & Handler, 2008, Raskin & Kircher, 2014). The scorers in our study were cut off from both quality control and algorithms. Our imposition of isolation may have created conditions quite different from the normal decision processes for some or most scorers.

What may be supportable from our findings are that some polygraph examiners who do not have, or do not avail themselves of, a means for independently evaluating their polygraph charts may be at risk from the influences of bias on their chart interpretation. The risk can vary depending on an individual susceptibility to biasing influences as not all examiners in our study were affected equally. Indeed, a small number showed total resistance to our manipulation while another small number showed the opposite. Our Bias Index suggests about half of our scorers showed no or low bias after our manipulation of expectations and the other were in moderate or high bias categories. It would be the latter two groups that would benefit most from independent evaluations of their polygraph data.

At a minimum, it would seem reasonable all examiners could refer to algorithm results to help protect their polygraph chart analysis from bias, as well as the eroding influences of fatigue, haste and distraction. More work is needed to determine the best role algorithms should play in decision-making. Algorithms can more efficiently process polygraph data than can humans and they are unaffected by biasing influences of outside information, perceptions and emotions. Nevertheless, algorithms so far do not appear ready to replace the human in the identification of certain artifacts, distortions and suspected countermeasures. This can be especially true if polygraph reactions have been induced by visual or auditory distractions that intrude upon the polygraph session, noticeable to the human but invisible to the algorithm.

It is our belief that neither the human nor the algorithm by themselves can compete with the optimal combination of human and algorithm (Raskin & Kircher, 1990). This remains largely conjecture for the moment. To date there has been little research interest in examiner-algorithm interaction in decision-making. The sole available study involving examiners having access to algorithm results as part of their evaluation process found examiners made almost no use of the algorithm information (Krapohl & Goodson, 2015). A combination of the algorithm and the human in that study improved decision accuracy more than 10%. Generalizability was limited, however, because it used Relevant-Irrelevant test data, an approach repeatedly found to have suspect validity (Honts & Handler, 2014; Horowitz, Kircher, Honts & Raskin, 1997; Krapohl & Goodson, 2015; Krapohl & Rosales, 2014). Research that includes techniques found valid and in common use would help resolve this question and guide field decision policies regarding algorithm use.

Independent quality control may also bring value as an alternative to, or in combination with, algorithms. Human evaluators kept blind to everything except the charts may be less affected by bias. Bias can still creep into decision-making, though. For example, if quality control reviewers overseeing screening examinations can infer the topics of the test questions from the question labels, they may be able to estimate the base rates of the different test questions. If so, they may unconsciously include this factor in the scoring of their polygraph charts.

Bias may have its greatest impact on techniques that do not have either a manual scoring system or an algorithm. The only technique still in wide practice at this writing that has no scoring system or algorithm is the Relevant-Irrelevant (RI) test. Evaluation of the RI test is still based on subjective impressions of the test charts. Lacking any benchmarks against which to conduct that evaluation, the RI test should be vulnerable to a range of individual differences among evaluators. Indeed, blind evaluation studies of confirmed RI screening cases where there was no bias manipulation found that pairs of examiners agreed about 59% of the time in their veracity decisions (Krapohl & Goodson, 2015; Krapohl & Rosales, 2014). The introduction of bias on top of poor inter-examiner agreement should be expected to produce a larger effect than what was found in the present study.

Despite the potential value of quality control and algorithms, the two established means to reduce examiner bias, they may not be immediately accepted by all members of the practitioner community. Many consider their own manual polygraph scores to be the final word on polygraph decisions. They may regard quality control and algorithms a loss of control, perhaps a diminishing of the examiner's value in the examination process. Independent quality control can add time and cost to polygraph examinations, the bane of examiners in private practice. As to algorithms, explaining to clients and customers why an algorithm disagrees with manual scoring may be uncomfortable to examiners who have become accustomed to delivering certainty with their scoring methods alone. These and other considerations must be overcome if there is to be an adaptive professional response to the influence of bias on polygraph results.

Limitations

Our study used a total of 15 cases, but only 5 in each category (truthful, deceptive, inconclusive). Whether all field cases are well-represented with this small sample seems unlikely. Our data may have been qualitatively different from the data polygraph examiners produce in their own sessions.

None of the 15 cases had been independently confirmed as coming from truthful or deceptive examinees. For this reason the study is not suitable for estimating polygraph decision accuracy. Any of the participating scorers may have been 100% accurate in his or her decisions.

The study had no control group, scorers who were given no biasing information. While we demonstrated significant differences between the groups, it remains possible that only one of the groups was affected by the manipulation. The other group may have matched the decisions made by a control group. While possible, this outcome seemed unlikely because an independent scorer had agreed with the original examiner for all 15 cases used in this study.

The Bias Index was intended to represent the degree to which expectations affected individual scorers. We acknowledge it is a blunt instrument that does not account for the influences of bias that are unrelated to the study manipulation. Consider a scorer with a personal bias for making decisions of truthfulness on polygraph cases who by chance was assigned to the truthful-expectation group. That scorer may receive a score in the Bias Index suggesting a bias for expectations but that score could have masked a pre-existing bias in the same direction. It may benefit future researchers to experimentally capture preexisting tendencies so to allow the researchers to disentangle the two potential sources of scorer bias.

Summary

Our study showed generally that polygraph examiners conducting blind scoring of polygraph cases without access to quality control or automated algorithms can fall prey to biasing influences. There was a range of individual differences in that susceptibility. The bias finding is consistent with research in almost all professions that rely entirely upon human judgments. The study, combined with those of previous research showing a similar effect, highlights the need for development of best practices and model policies to manage the potential problem of biasing influences on polygraph chart scoring.

References

American Psychological Association (2017). *Ethical Principles of Psychologists and Code of Conduct. Section 8.07: Deception in Research.* Available online at: https://www.apa.org/ethics/code/#807

Backster, C. (1969). *Technique Fundamentals of the Tri-Zone Polygraph Test.* Backster Research Foundation: New York.

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

Elaad, E., Ginton, A., and Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making, 7*(4) 279-292.

Honts, C.R., and Handler, M. (2014). Information gain in the Relevant-Irrelevant Test. Polygraph, 43(1), 137 – 147.

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

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 DoD Polygraph Institute, DoDPI02-R-0001.

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.

Krapohl, D.J, and Goodson, W. (2015). Decision accuracy for the Relevant-Irrelevant screening test: Influence of an algorithm on human decision-making. *European Polygraph, 9*(4), 189 – 208.

Krapohl, D.J. & Rosales, T. (2014). Decision accuracy for the Relevant-Irrelevant screening test: A partial replication. *Polygraph 43*(1), 20-29.

Krapohl, D.J., and Shaw, P.K. (2015). *Fundamentals of Polygraph Practice.* Academic Press: San Diego, CA.

Kubis, J. F. (1962). *Studies in Lie Detection: Computer Feasibility Considerations. RADC-TR 62-205, Contract AF 30(602)-2270.* Air Force Systems Command, U.S. Air Force, Griffiss Air Force Base. New York: Rome Air Development Center.

Larson, J.A. (1932). *Lying and its detection: A study of deception and deception tests.* University of Chicago Press: Chicago, IL.

Likert, R. (1932-3). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 5 – 55.

Matte, J.A. (1996). *Forensic Psychophysiology Using the Polygraph: Scientific Truth Verification – Lie Detection.* J.A.M Publications: Williamsville, NY.

Nelson, R. (2016). Scientific (analytic) theory of polygraph testing. *APA Magazine, 49(5),* 69-82.

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*(3), 185-215.

Raskin, D.C., Barland, G.H., and Podlesny, J.A. (1977). Validity and reliability of detection of deception. *Polygraph, 6*(1), 1-39.

Raskin, D. C. & Kircher J. C. (1990) Development of a computerized polygraph system and physiological measures for detection of deception and countermeasures: A pilot study. Unpublished manuscript.

Raskin, D. C., & Kircher, J. C. (2014). Validity of polygraph techniques and decision methods. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility assessment, scientific research and applications* (pp. 65–132). London: Academic Press.

Reid, J.E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law and Criminology, 34*, 542-547.

Shurany, T., Matte, J.A., and Stein, E. (2009). Influence of case facts on blind scorers of polygraph tests. *European Polygraph, 3* (3-4), 133 – 139.

Slowik, S., and Horvath, F. (2016). Book review: Fundamentals of Polygraph Practice. *European Polygraph, 10*(1), 27 – 36.

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

Winter, J.E. (1936). Comparison of the cardio-pneumo-psychograph and association methods in the detection of lying in cases of theft among college students. *Journal of Applied Psychology, 20*, 243 – 248.

Appendix A

Instructions Given to Scoring Volunteers

Please Read Before Beginning

Dear Scorer:

Thank you for volunteering to participate in our research project. Your data is important to this project. Our principal research objectives are to identify which polygraph channels are receiving the most points, how consistent the point assignment is within each channel, and how much agreement there is among polygraph data channels.

The 15 cases on the following pages were all confirmed by ground truth. Half of the examiners in this study will be given only truthful cases while others will have only received deceptive cases. You have been randomly selected to score truthful cases. The examinees were positively confirmed to be truthful to both relevant questions. However, the original examiners got some of these cases wrong. Here are additional details you will need to know:

1. All 15 cases were conducted using the Air Force Modified General Question Technique (AF MGQT) with two relevant questions.
2. Some cases are of suspects, others of possible witnesses or victims.
3. Some cases will be easy to score while others may be a bit of a challenge. The original examiners got most of these cases correct, but not all.
4. The topmost tracing is from the motion sensor, which was placed in the seat of the chair.
5. The bottommost tracing is the photoplethysmograph. Note: If you do not know how to score the photoplethysmograph, please leave those places empty in the score sheet.
6. Only assign one pneumo score per relevant question, though there are two pneumo tracings.
7. The remaining tracings are the regular polygraph channels placed in their traditional locations.
8. There are NO countermeasure cases in this sample.

We are asking that you use traditional 7-position scoring. Some volunteers normally use 3- position and the Empirical Scoring System (ESS) in their field practices, but we want all examiners to use 7-position scoring to make everyone's scores comparable.

We developed a score sheet for the cases in Excel. That file accompanied this document. We encourage volunteers to save the file on your computer, using your name in the File name (i.e., myname.xlsx). The score sheet automatically calculates the numbers we need for our study, and we hope you will find it easy to use. Once you score all 15 cases simply email the Excel file back to us at apakrapohl@gmail.com. If you cannot use the Excel score sheet, you are invited to use your own paper score sheets and just email back the total score for each data channel for each case. That means sending in one pneumo, one EDA, one cardio and one PPG total score per case. We will average your scores with those of other scorers to assess the weight each polygraph channel receives.

Though in this study we are not interested in decision accuracy, some volunteers expressed a wish to know how many cases they got right. Because you have already been told which scoring group you are in (truthful) you will be able to calculate your own accuracy. Based on our pilot study most scorers should get the majority of cases correct. Experienced examiners typically finish scoring all 15 cases in 1-2 hours, while less experienced examiners may take longer.

Finally, we ask scorers not to share the charts, their score sheets or their opinions about the study with anyone until after July 1st to protect the integrity of the data. July 1st is the final date for volunteers to submit their score sheets. After that date you are free to share anything you wish.

Once again, thank you for your support for this project. The results will be presented at the APA Seminar in August. If you have any questions, please write to Don at XXXXXX.

Good luck.

Appendix B

Email Sent to Scoring Volunteers

Dear APA Member:

Thank you for your interest in participating in the chart scoring project. We were gratified to receive so many positive responses to our request for scorers. We are currently preparing the 15 cases to be sent to you. They will be in PDF format attached to an email. If you do not have Adobe Reader, it can be downloaded for free at https://get.adobe.com/reader/.

We will ask you to enter your scores into an Excel spreadsheet which we will send you. If you do not have Excel, you may use paper score sheets, scan them, and send them to me at this email address as an attachment. Remember we are interested in your scores, not necessarily your decisions.

The most frequently asked question from the volunteers has been whether they will get to learn what ground truth is for these cases so they can know their accuracy in scoring. Because our interest in this study will focus on which polygraph channels get the most scores, we will provide ground truth when we send the cases. That way you can determine your own accuracy. Please be advised that the cases were confirmed by independent ground truth (e.g., confession, confession of someone else, fingerprints, determination that no crime took place, finding stolen property in the suspect's residence, etc.), and that the testing examiner may have been correct or incorrect in the original decision.

The second most frequent comment from volunteers has been that they know 7-position scoring but don't practice it often. For those who may be less practiced in 7-position scoring I have attached a pamphlet on the most commonly used form of 7-position scoring, the US government approach. I downloaded it from the Internet so I can't say it is exactly how the US government does scoring, but my impression is that it is similar to what I see examiners doing in the field most often. You are welcomed to use it as a reference for this project.

Finally, I need four pieces of background information from all scorers. As soon as possible please send me your:

1. Name
2. Whether you are an examiner in law enforcement, government or private practice
3. The city, state or province, and country where you do most of your testing
4. How many years of polygraph experience you have

Mr. Donnie Dutton and I will present the data from this project at the APA seminar in Austin, Texas, this August. We will not associate any scorings with any particular examiner. However, we intend to show a slide of the names of all of those individuals who stepped forward to volunteer for the project so we can publicly acknowledge and express appreciation for your help. If you do not wish to have your name listed on the Acknowledgments slide, please let me know and we will respect your wishes.

Again, thank you for volunteering to help us with this research project. We believe it will add to our growing knowledge about polygraph scoring.

Reducing Inconclusive Results: A Descriptive Analysis of Decision Rules, Weighted Electrodermal Scores and Multinomial Cut-scores

Raymond Nelson and Mark Handler[1]

Abstract

An archival sample of N=100 confirmed field polygraph exams was used to calculate descriptive statistics, including point estimates, to study the effect on inconclusive results and other metrics of test accuracy. Data were analyzed as a function of different decision rules, structural weighting of sensor scores and cut-scores. Decision rules included the federal-zone-rules, grand-total-rule, subtotal-score-rule, two-stage-rules. Numerical scores were obtained through an automated feature extraction algorithm. Results were evaluated with both unweighted three-position numerical scores and after weighting the integer scores for electrodermal responses. Results are shown for both traditional numerical cut-scores and also using cut-scores obtained from multinomial reference distributions for three-position comparison question test scores with both weighted and unweighted electrodermal scores. Use of different decision rules had less effect with the multinomial cut-scores than with traditional cut-scores. Weighted EDA scores produced an average 49% reduction of inconclusive results across all decision rules, and the combination of weighted EDA scores and multinomial cut-scores reduced the occurrence of inconclusive results by an average of 72%.

Introduction

Few things are more disappointing for field polygraph examiners and referring agents than an inconclusive[2] test result. There is little to compare with the sense of frustration when – after all the time and effort invested in preparing for the test, the pretest interview, target selection, question formulation, test data collection, and test data analysis – a test result is not statistically significant for deception or truth-telling. This sense of frustration is, at times, shared by some examinees – especially those who are innocent – who, having agreed to testing in hopes of producing test data that can serve as a basis of evidence to support professional conclusions about innocence or truth-telling, must necessarily be informed that their inconclusive test result provides no better information about truth or deception than was already available before the test.

In years past, in the absence of a probabilistic view of polygraph test results, it may have been tempting to up-sell the capabilities of the polygraph as virtually infallible. During that era, an inconclusive result was at times regarded as an indication of an unskilled examiner. Behind this attitude or belief was likely a sincere desire among polygraph examiners to be of actual help to referring professionals, for whom an inconclusive test result offered little or no practical value. And virtually nobody wants to purchase the services of a probabilistic test for which we could offer little more than conjecture as to the strength of the conclusion. In the absence of an ability to realistically quantify the level of confidence

1 The authors are extremely grateful to Mr. Don Krapohl and Texas DPS Captain Matt Hicks (who functioned as the action editor for this project) for reviewing, commenting and editing earlier drafts of this manuscript.

2 To remain consistent with the terminology of the forensic sciences we have chosen to call indeterminate polygraph results “inconclusive” rather than the neologism “No Opinion.”

or margin of uncertainty for the test result, it may have been a matter of professional marketing to up-sell or over-sell the capabilities of the test, including unrealistic expectations for infallibility and the absence of inconclusive results.

It may be helpful to remember that inconclusive results are not unique to polygraph testing; all forensic tests are burdened with a certain proportion of inconclusive outcomes. The stated reason for them depends upon the forensic discipline in which the test is conducted, but in the main it is because the signal/pattern/trace/sample/marker/image is inadequate or contaminated. So too is it with the polygraph. Even under ideal conditions, with heroic effort and with perfect examinees, pretest interviews, and charts there will always be some cases in which the data do not allow for a reliable decision. An important objective in all forensic disciplines is to minimize those occasions.

Objectively, if people believed the polygraph to be infallible then there would be no great difficulty in accepting a test result at face value. However, most polygraph examiners, and other professionals that we are aware of, would be hesitant to accept a test result at face value without scrutinizing the test administration, test data, and analytic result. Also, no thinking person today will accept the notion that polygraph is infallible or deterministic. It is well established that all scientific test results rely on probabilities and probability theory to quantify important phenomena that cannot be subject to perfect deterministic observation or direct physical measurement. Polygraph technologies, methodologies, and standards have all evolved along with this understanding.

Presently, it is reasonably understood by most – polygraph professionals, referring agents, courts, legislators, scientists, members of the community, and the entertainment and news industries – that the polygraph test is, like other scientific tests, merely a statistical classifier intended to quantify a phenomenon that cannot be subject to perfect deterministic observation or direct physical measurement. All tests are fundamentally probabilistic. Along with any use of statistics and probability theory comes the potential for testing error. The practical purpose of test data analysis is often to achieve a categorical classification or categorical test result. But experts who possess a broader and more complete understanding of scientific test are aware that the actual purpose of test data analysis is to quantify – in some reproducible manner – the level of confidence or margin of uncertainty surrounding a test result. Any reasonable and intelligent use of statistical theory and statistical methods will include an acknowledgement of some potential for testing error, and some potential that test results may not achieve a required level of statistical significance. Despite this, inclusive test results, are still the bane of field polygraph examiners everywhere.

Reducing the proportion of inconclusive results can potentially increase the effectiveness of the test. Inconclusive results are inversely related to utility. Higher rates of inconclusive results correspond with a lower rate of utility or usefulness. One common strategy for reducing the occurrence of inconclusive test results is to increase the volume of available test data by repeating the series of relevant and comparison questions one or two additional times, beyond the minimum three repetitions. Another strategy for reducing the occurrence of inconclusive results is to include additional independent diagnostic information, such as from a vasomotor sensor, to the traditional array of respiration, cardio, and electrodermal sensors. There are also other, subtler, strategies to reduce the occurrence of inconclusive test results, and these can include target selection, question formulation, and interviewing skills.

In this paper we use an archival sample of confirmed criminal investigation polygraphs to show three additional evidence-based approaches to field practices that can reduce the occurrence of inconclusive test results while increasing the objectivity and statistical power of the test. The first among these involves *polygraph decision rules*. The second field practice area involves the *weighting of electrodermal* (EDA) scores when assigning Likert-type (1932) three-position scores to physiological response to polygraph questions. Thirdly, we show the effect of *replacing traditional polygraph cut-scores* with cut-scores derived from a multinomial distribution of polygraph scores based the analytic theory of the polygraph test.

Method

Sample Data

A sample of confirmed field polygraphs was obtained from an earlier study by Krapohl (2005). The sample data consisted of N=100 polygraph examinations that were selected randomly from an archive of confirmed field cases. All exams were conducted by U.S. law enforcement agencies using the Federal Zone Comparison (ZCT) test format (Light, 1999; Department of Defense, 2006a), for which the sequence of scored questions, including three relevant questions and three comparison questions, was repeated three times. Sample cases consisted of n=50 confirmed truthful cases and n=50 confirmed deceptive cases. Data for all cases consisted of recording sensors for changes in thoracic and abdominal respiration activity, electrodermal activity, and cardiovascular activity.

Table 1 shows is an extended calculation of the sample results reported by Krapohl (2005) using the Federal Zone Rule (FZR) and seven-position scores, including test sensitivity or true-positive (TP) rate and specificity or true-negative (TN) rate, false-positive (FP) and false-negative (FN) errors, inconclusive (INC) results, and the unweighted average of correct decisions and inconclusive results for confirmed deceptive and truthful cases. Table 1 also shows the positive-predictive-value (PPV; calculated as TP/(TP+FP) and negative-predictive-value (NPV; calculated as TN/(TN+FN). Also shown in Table 1 is the detection efficiency coefficient (DEC; Kircher, Horowitz & Raskin, 1988), calculated as the Pearson correlation of the case status [-1, 1] and test result [-1, 0, 1]. The DEC provides a single metric that encompasses correct classifications, errors and inconclusive results for both deceptive and truthful sample cases.

Krapohl (2005) reported results using "investigative rules," for which deceptive classifications were made if the grand total score equaled or exceeded a cut-score of -6, or any subtotal score equaled or exceeded -3. Truthful classifications were made only when the grand-total score equaled or exceeded +6 *and* all subtotal scores exceed +1. All other conditions were classified as inconclusive. These procedures can be found in publications by the Department of Defense (2006a; 2006b) and is referred to as the Federal Zone Rule (FZR) for the remainder of this manuscript.

Krapohl (2005) also reported results using "evidentiary rules," for which deceptive classifications were made if the grand total score equaled or exceeded -6 while truthful classifications were made if the grand total score equaled or exceeded +4. In cases where the grand total was in the range from -5 to +3 the subtotal scores were used to make deceptive classifications if any subtotal equaled or exceeded -3. All other conditions were classified as inconclusive. The process of first using the grand total score and subsequently using subsequently using subtotal scores if the grand total is inconclusive was first described by Senter and Dollins (2003) and is referred to as the two-stage-rule (TSR) for the remainder of this manuscript.

Table 1. Sample results (n=100) reported by Krapohl (2005) for seven-position scores.

	Investigative Rules / FZR	Evidentiary Rules / TSR
Sensitivity (deception)	.78	.81
Specificity (truth-telling)	.62	.80
False-negative Errors	.07	.10
False-positive Errors	.09	.09
Inconclusive-guilty	.15	.09
Inconclusive-innocent	.29	.11
Unweighted Accuracy	.90	.90
Unweighted Inconclusives	.22	.10
Positive Predictive Value	.90	.90
Negative Predictive Value	.90	.89
Detection Efficiency Coefficient	.67	.74

Results from Krapohl (2005), shown in Table 1, results are consistent with other studies using the FZR and showed that inconclusive results are loaded for innocent persons. Krapohl showed that use of the seven-position scores with the TSR and evidentiary cut-scores resulted in similar classification accuracy, though with improved test specificity and a reduction of the inconclusive rate by 55%. Importantly, Krapohl used different cut-scores for the FZR and TSR, and it remains unknown what portion of the observed difference can be attributed to decision rules and/or to the different cut-scores. The present analysis is an attempt to provide more information about observed differences in test accuracy as a function of inconclusive rates for decision rules, cut-scores and other factors.

Feature Extraction and Data Reduction

Sample case were scored using a three-position numerical scoring method (Bradley & Janisse, 1981; Department of Defense, 2006b; van Herk, 1991). For each iteration of each relevant question and for each recording sensor, three-position numerical scores were assigned via an automated feature extraction algorithm that was developed using the R statistical computing language (R Core Team, 2018). The three-position scoring method is Likert-type coding system, based on the analytic theory of the polygraph test (Nelson, 2015). Scores of +1, 0 and -1 were assigned to relevant and comparison question pairs – referred to by field polygraph examiners as *analysis subtotals or "spots".* Negative scores were indicative of greater changes in physiological activity in response to relevant questions whereas positive scores were indicative of greater changes in physiological activity in response to comparison questions. Scores of 0 were assigned when there was little or no difference in response to relevant and comparison questions.

The automated algorithm was designed to extract information about the relative amplitude of increase in electrodermal activity, relative increase in blood pressure, and relative suppression or reduction of respiration activity. These responses have been shown to be correlated with differences in physiological response to relevant and comparison polygraph questions under the analytic theory of the comparison question polygraph test (Kircher & Raskin, 1988; Kubis, 1962; Summers, 1939) [see Nelson (2016) for a discussion]. The automated feature extraction algorithm performed nearly all traditional analysis tasks, including identification of response onset and response peak and the calculation of numerical differences between response peak and response onset.

The automated feature extraction algorithm also selected relevant and comparison question pairs, such that for each repetition of the question sequences for the ZCT cases in the sample data the physiological responses to second and third relevant questions was paired with the physiological response to the preceding comparison question. For each recording sensor the automated algorithm paired the physiological responses to the first relevant question with the physiological response from either the preceding or subsequent comparison question by selecting the comparison question that produced the greater change in physiological activity.

Prior to feature extraction the automated algorithm scaled the recorded physiological data for visualization and performed some identification and rejection of data artifacts such as deep breaths in respiration activity, physical movement in the cardio data, and labile electrodermal responses that were unrelated to or untimely with the test question.

Dimensionless numerical measurements were obtained from the scaled physiological data from response onset to response end using a 15 second evaluation window (EW). Respiration excursion was measured as the mean of 1 second intervals from 0 to 14 seconds at a data rate of 30 sample per second, excluding 2 seconds prior to and subsequent to the verbal answer. Electrodermal and cardio responses were measured as the dimensionless difference of the maximum difference between the onset of a positive slope segment that began during the ROW and a subsequent peak of a positive slope segment. EDA and cardio response peak points were identified in the EW and were included in the extracted measurements if they occurred after the EW if the positive slope segment began during the ROW.

Measured segments for relevant and comparison questions were combined as the R/C ratio, which is the extracted relevant question measurement divided by the extracted comparison question measurement. R/C ratios were logged so that they produce values that are symmetrical around a mean of zero.

Non-parametric integer scores were assigned to the physiological responses to analysis spots (i.e., relevant and comparison question pairs) using a three-position Likert (1932) type scale [+1, 0 -1]. Integer scores of positive sign value were assigned when the change in physiology was greater at the selected comparison question than the relevant question. Integer scores of negative sign value were assigned when the change in physiology was greater at the relevant question than the selected comparison question. Scores of zero sign (numerical zero) were assigned when the data were distorted from movement/activity artifact or were insufficient for scoring due to no response or due to a response onset prior to question onset, with no response onset during a defined response onset window (ROW) from question onset to five seconds after the verbal answer. Electrodermal and cardio responses were not used if they began during a .5 second latency period at question onset.

Non-parametric scores are intended to capture and express general information about the differences in response to relevant and comparison questions but do not provide information that can be subject to linear assumptions about the degree of difference in response. However, some threshold constraints were used to prevent the assignment of numerical scores for segments of data that produced very little difference or extreme differences in responses to relevant and comparison questions. Threshold constraints were determined from optimization experiments with another a confirmed case sample used by Nelson (2018a) and are shown in Table 2. Integer scores were assigned when the logged R/C ratio was within these constraints, and no score was assigned with the value was less than or greater than these constraint thresholds.

Table 2. Threshold constraints for non-parametric scores

-	Lower limit		Upper limit	
Sensor	Ratio	Logged ratio	Ratio	Logged ratio
Respiration	1.05/1.25*	.049/.223	1.5	.406
Electrodermal	1.05	.049	1000	6.908
Cardio	1.05	.049	1000	6.908

* An asymmetrical constraint 1.25 was used for the lower limit of respiration scores when assigning scores of + sign value. Optimization studies with other sampling data indicates that + scores are more likely to be negatively correlated with truth-telling without the asymmetrical constraint.

Data Reduction

Data reduction for each of the sample cases was accomplished by summing the numerical scores for respiration, electrodermal and cardio sensors for all presentations of each of the relevant test questions. In this way, a numerical subtotal score was obtained for each of the relevant questions. Subtotal scores were then summed to obtain a grand total score for each of the sample cases. Classifications of deception and truth-telling would be made with grand total and/or subtotal scores according to established polygraph decision rules.

Analysis

Sample cases were analyzed for correct decisions, errors and inconclusive results for the confirmed field cases and were calculated for test sensitivity, specificity, false-negative and false-positive errors, and inconclusive results for guilty and innocent sub-groups. In addition, the proportion of correct decisions was calculated for the guilty and innocent cases after excluding inconclusive results, along with the unweighted average of decision accuracy for the two groups. Positive predictive value and negative predictive values were also calculated as well as Detection efficiency coefficients. Accuracy indices were calculated for several conditions, including for three-position scores using traditional numerical cut-scores, after weighting the EDA scores, and using cut-scores selected from a multinomial reference distribution of weighted and unweighted three-position scores (Nelson, 2017, Nelson, 2018b).

Results

Results were tabulated for three-position scores using different decision rules and traditional numerical cut-scores. Results were also tabulated after doubling the value of all electrodermal scores. Finally, results were tabulated using cut-scores that were selected from multinomial reference distributions for both weighted three-position and unweighted EDA scores.

Decision Rules

Previous studies by Senter & Dollins (2003) have suggested that the choice of decision rules may play an important role in the effectiveness of polygraph classifications of deception and truth-telling. [See Nelson (2018c) for a discussion of different decision rules]. Table 3 shows the results three-position numerical scores for the n=100 confirmed field cases. Results are shown using the Federal ZCT Rule (FZR) with traditional numerical cut-scores (described earlier). Results for this study were also calculated using the Grand-Total Rule (GTR). Use of the GTR is a matter of summing all numerical scores and comparing the result to numerical cut-scores for deception or truth-telling (traditionally +6 and -6). Results were also calculated using the subtotal score rule (SSR), for which the lowest question subtotal (strongest indication of deception) is compared to numerical cut-scores for truth-telling (traditionally +3 or greater for all subtotals) or deception (traditionally -3 or lower for any subtotal). Also shown in Table 3 are of the sample cases using automated feature extraction and score assignment using the TSR with both traditional numerical cut-scores and the evidentiary cut-scores proposed by Krapohl (2005), as described earlier.

Table 3. Sample results for decision rules with three-position scores and traditional cut-scores (n=100).

	FZR	GTR	SSR	TSR*	EDR/TSR**
Sensitivity (deception)	.80	.56	.80	.80	.80
Specificity (truth-telling)	.32	.32	.06	.32	.48
False-negative Errors	<.01	<.01	<.01	<.01	<.01
False-positive Errors	.10	<.01	.10	.10	.10
Inconclusive-guilty	.20	.44	.20	.20	.20
Inconclusive-innocent	.58	.68	.84	.58	.42
Unweighted Accuracy	.88	>.99	.69	.88	.91
Unweighted Inconclusive rate	.39	.56	.52	.39	.31
Positive Predictive Value	.89	>.99	.89	.89	.89
Negative Predictive Value	>.99	>.99	.86	>.99	>.99
Detection Efficiency Coefficient	.79	.67	.79	.76	.82

* Results with the TSR are shown using the traditional numerical cut-scores.

** For comparison with Table 1, these evidentiary decision rules (EDR) results with the TSR are shown using asymmetrical cut-scores using by Krapohl (2005).

Test sensitivity in Table 3 is similar to that in Table 1, though specificity rates are lower for all decision rules. PPV and NPV for the three-position scores were greater than the results in Table 1, along with generally higher rates of inconclusive results. FN errors for the automated three-position scores was lower than in Table1. DEC was also increased for the FZR and TSR, and this can be attributed to the observed reduction in FN errors. It possible that these differences are due to the use of three-position vs seven-position numerical scores, though it is unknown what differences may be due to the use of automated feature extraction vs visual/subjective feature extraction.

PPV and NPV were highest for the GTR, though the inconclusive rate was also greatest for this decision rule. Light (1999) argued the rate of inconclusive cases for the GTR was unacceptable for law enforcement use. However, the Light study is limited in scope – involving only confirmed guilty cases – and Table 3 shows that inconclusive rates are loaded on innocent cases. Also, Light did not include an evaluation of FP errors or the effectiveness of different numerical cut-scores. Inconclusive results for guilty cases in Table 3 are greater than Table 1, and this is most likely attributed to differences in seven-position and three-position scores.

In the present study, using three-position scores and automated feature extraction, it is unclear whether the TSR provides any real advantage – in terms of classification accuracy – over the FZR. Results shown in Table 3 indicate a reduction of inconclusive results for innocent cases as a result of improved numerical cut-scores with the TSR. DEC was greatest for the TSR with the improved numerical cut-scores. This suggests the possibility that traditional numerical cut-scores for grand-total scores are inefficient for the reduced three-position scale.

Weighted Electrodermal Scores

Several publications have suggested that EDA data account for a larger portion of the diagnostic variance in test scores (Ansley & Krapohl, 2000; Harris, Horner & McQuarrie, 2000; Harris & Olson, 1994; Kircher, 1981, 1983; Kircher, Kristjianson, Gardner & Webb, 2005; Kircher & Raskin, 1988; Krapohl & McManus, 1999; Nelson, Krapohl & Handler, 2008; Raskin, Kircher, Honts & Horowitz, 1988) and contribute more information to effective conclusions about deception or truth-telling compared to the other recording sensors. To observe the differences in sample results with the three-position scores, EDA scores were doubled in value in the manner previously described by Krapohl and McManus (1999). Table 4 shows the sample results after weighting the electrodermal scores.

Table 4. Sample results for decision rules with n=100 confirmed field cases with traditional cut-scores and weighted electrodermal scores.

	FZR	GTR	SSR	TSR*	EDR/TSR**
Sensitivity (deception)	.92	.66	.92	.92	.92
Specificity (truth-telling)	.52	.54	.16	.54	.68
False-negative Errors	<.01	<.01	<.01	<.01	.02
False-positive Errors	.24	<.01	.24	.22	.14
Inconclusive-guilty	.08	.34	.08	.08	.06
Inconclusive-innocent	.24	.46	.60	.24	.18
Unweighted Accuracy	.84	>.99	.70	.86	.90
Unweighted Inconclusives	.16	.40	.34	.16	.12
Positive Predictive Value	.79	>.99	.79	.81	.87
Negative Predictive Value	>.99	>.99	>.99	>.99	.97
Detection Efficiency Coefficient	.87	.78	.86	.88	.90

* Results with the TSR are shown using the traditional numerical cut-scores.

** For comparison with Table 1, these evidentiary decision rules (EDR) results with the TSR are shown using asymmetrical cut-scores using by Krapohl (2005).

Weighting the EDA scores more than the other sensor score increased test sensitivity and specificity and reduced the occurrence of inconclusive results by 59% for the FZR, 29% for the GTR, and 35% for the SSR. The reduction of inconclusive results was 59% for the TSR with traditional cut-scores and 61% for the TSR using cut-scores that were suggested as optimal for evidentiary exams. The reduction of inconclusive results was greater for innocent cases than for guilty cases for all decision rules.

The GTR provided the greatest overall classification accuracy, though FN errors were low for all decision rules. The most obvious effect from weighting the EDA scores in Table 4, compared to Table 3, was a reduction of inconclusive results, along with increases in both test sensitivity to deception and specificity to truth-telling. The reduction of inconclusive results was greatest for the confirmed innocent cases. Interestingly, PPV was reduced for all models except the GTR which produced lower FN and FP error rates than other decision rules. DEC was improved for all decision rules and was greatest for the TSR with improved cut-scores. This suggests that the selection of cut-scores may be an important consideration in the management and reduction of classification errors or inconclusive results.

Multinomial Cut-Scores

Multinomial Cut-Scores with Weighted Three-Position Scores

Weighted three-position scores were evaluated using cut-scores selected from multinomial reference distributions described by Nelson (2017). [See Nelson (2018b) for a discussion of how to use the multinomial reference distributions.] Multinomial cut-scores for weighted three-position of event-specific polygraphs with three relevant questions, were as follows: grand total >= +3 for truthful classifications or <= -3 for deceptive classifications. Deceptive classifications were made using the subtotal scores when the grand total score was inconclusive if any subtotal score <= -7. The cut-score for subtotal scores was determined using a statistical correction for multiplicity effects to avoid the potential inflation of FP errors when using multiple subtotal scores for deceptive classifications. Results are shown in Table 5 for the combination of weighted EDA scores and multinomial cut-scores.

Table 5. Sample results (n=100) using weighted three-position scores with multinomial cut-scores.

	FZR	GTR	SSR	TSR
Sensitivity (deception)	.92	.88	.92	.92
Specificity (truth-telling)	.80	.80	.34	.80
False-negative Errors	.04	.04	.02	.04
False-positive Errors	.08	.06	.24	.08
Inconclusive-guilty	.04	.08	.06	.04
Inconclusive-innocent	.12	.14	.42	.12
Unweighted Accuracy	.93	.94	.78	.93
Unweighted Inconclusives	.08	.11	.24	.08
Positive Predictive Value	.92	.94	.79	.92
Negative Predictive Value	.95	.95	.94	.95
Detection Efficiency Coefficient	.94	.94	.85	.94

Use of multinomial cut-scores further improved test sensitivity to deception for the GTR, and improved test specificity to truth-telling for the FZR, GTR, SSR and TSR. Compared to the use of weighted electrodermal scores with traditional numerical cut-scores the multinomial cut-scores produced a reduction in inconclusive results by 50% for the FZR, 73% for the GTR, 29% for the SSR and 33% for the TSR. DECs for the multinomial cut-scores were consistently greater than for the traditional numerical cut-scores.

Not surprisingly, the SSR showed high sensitivity to deception, though not greater than any of the other decision rules, along with weaker specificity to truth-telling and a higher inconclusive rate that was loaded on innocent cases. Overall decision accuracy and DEC for the SSR was lower and inconclusive rates were higher than for other decision rules that included the use of the grand total score. This difference may be attributable to inherent multiplicity when using subtotal scores, and also to the smaller volume of information available to support decisions based on individual subtotal scores.

Multinomial Cut-Scores with Unweighted Three-Position Scores

Three-position scores were also evaluated using cut-scores obtained from a multinomial reference distribution for polygraph exams with three relevant questions and three to five charts. Multinomial cut-scores for the unweighted three-position scores were as follows: grand-total >= +2 for truthful classifications or <= -2 for deceptive classifications. Deceptive classifications were made using the subtotal scores when the grand-total score was inconclusive if any subtotal score <= -6. The cut-score for subtotal scores was determined using a statistical correction for multiplicity effects to avoid the potential inflation of FP errors when using multiple subtotal scores for deceptive classifications. Results are shown in Table 6 for the combination of weighted EDA scores and multinomial cut-scores.

Table 6. Sample results (n=100) using unweighted three-position scores with multinomial cut-scores.

	FZR	GTR	SSR	TSR
Sensitivity (deception)	.88	.86	.80	.88
Specificity (truth-telling)	.78	.76	.30	.78
False-negative Errors	.04	.04	.02	.04
False-positive Errors	.04	.04	.10	.04
Inconclusive-guilty	.08	.10	.18	.08
Inconclusive-innocent	.20	.20	.60	.20
Unweighted Accuracy	.95	.95	.86	.95
Unweighted Inconclusives	.14	.15	.39	.14
Positive Predictive Value	.96	.96	.89	.96
Negative Predictive Value	.95	.95	.94	.95
Detection Efficiency Coefficient	.91	.90	.79	.91

Use of multinomial cut-scores improved the effectiveness of classifications with the unweighted three-position scores. Classification accuracy with the three-position scores using multinomial cut-scores was similar to that of the weighted multinomial model. However, both sensitivity and specificity were reduced slightly for the unweighted three-position scores. Compared to the weighted model the inconclusive rate for unweighted three-position scores increased by an average of 62% for all decision rules. The increase in inconclusive results was loaded on innocent cases. The unweighted three position scores also produced fewer false-positive errors. These differences can appear substantial when described as a percentage of change. The SSR underperformed relative to the other decision rules, with weaker test specificity and higher inconclusive results. Overall detection accuracy for the multinomial three-position model was high, though did not equal the effectiveness of the weighted three-position model.

Summary

This project involved the calculation of descriptive statistics for test accuracy, error and inconclusive rates as a function of different decision rules, structural weighting of sensor scores, and cut-scores. Results show that a number of procedural and field practice decision can have an important impact on the criterion accuracy of polygraph test results. Although PPV and NPV were consistently high for most conditions, differences in the rate of inconclusive results can be observed and this can directly affect the power of the test in terms of test sensitivity, specificity and error rates.

Results for the three-position scores, shown in Table 3, were similar for the FZR and TSR. However, there was a small reduction of inconclusive results, along with a corresponding increase in test specificity when using the TSR with cut-scores that were suggested as more optimal for evidentiary polygraph testing. Weighting the EDA scores, shown in Table 4, more than the other sensor scores increased test sensitivity and specificity, and reduced the occurrence of inconclusive results and average of 49%, compared to results from unweighted three-position scores, for all decision rules. Use of multinomial cut-scores produced further reductions in the occurrence of inconclusive results along with further increases in DECs for all decision rules.

The combination of weighted EDA scores and multinomial cut-scores reduced the occurrence of inconclusive results by an average of 72% across all decision rules, compared to unweighted three-position scores and traditional numerical cut-scores. The reduction in inconclusive results was greatest for the GTR, which does not make use of subtotal scores, and for which the rate of inconclusive results with multinomial cut-scores was closer to that of other decision rules.

Field practitioners have provided anecdotal information suggesting that their observed rates of inconclusive rates are inconsistent with, and lower than, those in published studies. This is understandable because field practitioners, working at the level of individual cases, may be ethically justified in engaging in practices intended to resolve or reduce the occurrence of inconclusive results (e.g., conducting additional repetitions of the question sequence, or repeating an examination). In contrast researchers who work with samples of cases would be vulnerable to suggestions of manipulating a research outcome if they were to engage in such actions at level of some, though not all, individual cases. The result is that inconclusive rates in field practice may continue to be lower than those reported in published studies.

The most effective model in this analysis – illustrated by sensitivity, specificity, inconclusive and DECs in Table 5 – was with the TSR using weighted EDA using multinomial cut-scores. Interestingly, accuracies for the FZR, and TSR were effectively identical for the weighted three-position scores with multinomial cut-scores, suggesting that the selection of cut-scores may be more important than the decision rules. Similarity of the DEC for the GTR, FZR and TSR provide further indication of this, and suggest that some previously reported conclusions about the GTR may have been unduly influenced by reliance on traditional numerical cut-scores that were inefficient for grand total scores.

This project, like all projects, is not without some limitations. The most obvious limitation is the small sample size (N=100). Though moderately sized for a project of this type, it is axiomatic that larger sample sizes are more easily viewed as comfortably approximating the population. However, sample size is not the only, or primary, consideration when attempting to understand the representativeness of a sample – for which random selection may be more important. This project, involving an archival sample, is necessarily precluded from any influence due to sampling methodology that is not presently expressed in the sampling data. It is also, necessarily dependent upon assumptions that the sampling data are in some way informative.

Another important limitation of this project is that no tests of statistical significance were completed. This was by design, as it was hoped that a descriptive approach to the statistical analysis might be of greater practical value to polygraph field examiners and program managers who may be more familiar and conversant with field practice policy decisions than multiple ANOVA. Future research should include a more complete analysis of the variance of the related effect sizes for polygraph decision rules, structural weighting coefficients for sensor scores, and cut-scores. Also, no statistical confidence intervals were included in this document, though informed readers can easily use a number of methods to calculate the confidence intervals of interest.

This project involved only a field sample of confirmed criminal investigation (event-specific) polygraphs and did not include a sample of multiple issue screening polygraphs. We suggest that some cautious generalization of these results is still in order. This is because of practical and important difference between event-specific diagnostic polygraphs and screening polygraphs that involve assumptions about the independence of multiple-issue screening questions. These assumptions are at best *convenience assumptions* because they assume independence in that different test items have no shared source of response variance – that whatever could influence responses to each item could have not affected any other item. As it happens, all polygraph questions within any examination will always have some shared source of response variance – in the form of the attention of the examinee. Both event-specific diagnostic and multiple-issue screening polygraphs will also be influenced by statistical multiplicity effects as determined by the selection of polygraph decision rules. For these reasons, a similar pattern of results can be expected for polygraph screening exams as is observed with this sample data.

Astute readers will note that this project does not attempt to discuss all possible methods of reducing inconclusive rates and improving polygraph test effectiveness. Some of those other methods may include: interviewing approaches, quality assurance activities, greater use of automation, use of the vasomotor sensor, recording additional charts, clarification of operational definitions, use of interview route-maps, refined target selection, and/or improvements in question formulation. All of these should remain as areas for continued research and development.

Results from this study point clearly to the fact that traditional numerical cut-scores are effective at producing a low FN rate but are burdened with un-necessarily weak test specificity to truth-telling and un-necessarily high rates of inconclusive results. An interesting observation that can be made about these results is that there was no advantage to the use of the SSR in terms of increased test sensitivity to deception, when compared to the other decision rules. The FN rate for the SSR was equal to that of the other decision rules for when using traditional cut-scores and was reduced from that of the TSR and FZR by 50% (.02 / .04), while the FP rate increased by a factor of 3 (.24 / .08) for the SSR. The inconclusive rate was larger for the SSR than for the FZR and TSR and was loaded on innocent cases. As shown in Table 3, the SSR was especially weak with unweighted three-position scores and traditional numerical cut-scores. The practical implication of these observations is that it may be difficult to justify the use of the SSR outside of polygraph screening contexts, where it some over-prediction may be desirable or intended – and difficult to justify the use of the SSR when using unweighted score and traditional cut-scores.

These results show clearly that optimization of field practices in each of these areas – decision rules, weighting of EDA scores and the selection of cut-scores – can provide important advantages to many, including polygraph field examiners, program managers, courts, legislators, researchers, and polygraph examinees. Further exploration is needed to better understand the utility functions in terms of economic values and operational costs associated with sensitivity, specificity, FN and FP errors, and inconclusive results rates. Greater reliance on statistical measurement theory can permit polygraph programs to refine their policies to better achieve their mission objectives and goals.

Inconclusive test results are likely to persist as a bane to polygraph field examiners, program managers, and others – including polygraph examinees. The availability of evidence-based procedural solutions that can reliably reduce the occurrence of inconclusive test results appears to be worthy of further attention and consideration.

References

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

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

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].

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.

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

Kircher, J. C. (1981). *Computerized chart evaluation in the detection of deception.* University of Utah. [Thesis],

Kircher, J. C. (1983). *Computerized decision making and patterns of activation in the detection of deception.* Doctoral dissertation, University of Utah, Salt Lake City. Dissertation Abstracts International, 44, 345.

Kircher, J. C., Horowitz, S. W. & Raskin, D.C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior, 12*, 79-90.

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.

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

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

Kubis, J. F. (1962). *Studies in Lie Detection: Computer Feasibility Considerations. RADC-TR 62-205, Contract AF 30(602)-2270.* Air Force Systems Command, U.S. Air Force, Griffiss Air Force Base. New York: Rome Air Development Center.

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

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

Nelson, R. (2016). Scientific (analytic) theory of polygraph testing. *APA Magazine, 49(5),* 69-82.

Nelson, R. (2017). Multinomial reference distributions for the Empirical Scoring System. *Polygraph & Forensic Credibility Assessment,46(2),* 81-115.

Nelson, R. (2018a). Electrodermal signal processing: A correlation study of auto-centered EDA and manually-centered EDA with the criterion state of deception and truth-telling. *Polygraph & Forensic Credibility Assessment, 47 (1)*, 53-65.

Nelson, R. (2018b). Guide for how to use the ESS-multinomial reference tables in four steps. *APA Magazine, 51(2),* 78-89.

Nelson, R. (2018c). Practical polygraph: A survey and description of decision rules. *APA Magazine, 51(2),* 127-133.

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.

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.

Senter, S. M. & Dollins, A.B. (2003). *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. Reprinted in Polygraph, 37(2), 149-164.

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

Van Herk, M. (1990). Numerical evaluation: Seven point scale +/-6 and possible alternatives: A discussion. *The Newsletter of the Canadian Association of Police Polygraphists, 7,* 28-47. Reprinted in Polygraph, 20(2), 70-79.

Practical Polygraph: The Known-Solution Acquaintance Test and Functionality Check

Raymond Nelson, Rodolfo Prado, Ben Blalock and Mark Handler

Abstract

We discuss the published literature on the effectiveness of the known-solution acquaintance test (ACQT). Although studies have not supported an effect for the unknown-solution ACQT, the known-solution ACQT is better supported by scientific evidence. Some important advantages of the known-solution ACQT is that it does not engage a role-reversal, wherein the examinee is testing the polygraph test, and does not rely on manipulation or deception as a form of demonstration of the validity of the polygraph testing. The basic procedures are described for using the known-solution ACQT.

Introduction

Use of an acquaintance test (ACQT) is a standard practice for polygraph diagnostic and screening exams (American Polygraph Association, 2018; Department of Defense, 2006), and serves as a form of *practice-test* for the examinee and as a *functionality-check*[1] for the examiner. Reid (1947) provided an early description of field practices that were in some ways like the contemporary ACQT, involving a *card-test* as a form of *control-test* to ascertain that the examinee can exhibit normal physiological responses to the test questions in the event of deception. In contemporary usage, the ACQT is conducted as the first test chart. However, early use of the *card-control* test was after the first iteration of the sequence of polygraph test questions and its purported purpose was to demonstrate the polygraph effectiveness to the examinee.

Discussion

Reid and Inbau (1966) described the card-control test wherein the examinee is instructed to pick a card under the false pretext that the examiner does not know which card was chosen. Although Summers (1939) first described the use of comparative response questions, Reid (1947) popularized the idea and suggested that the inclusion of control questions in polygraph test question sequencing represented an advancement in polygraph field practice methodology over the card-control test. Reid also suggested that the card-control test should be conducted as the first chart. Other publications, such as those by Kirby (1981), describe continued use of the standard-card-test, also known as a *card-stim* test, after the first iteration of the polygraph test questions. Later, Widup and Barland (1994) reported there was no effect for the classification of deception or truth-telling, and no effect for inconclusive results, when a *number-stim* test was used before or after the first iteration of the sequence of polygraph test questions.

Kirby (1981) studied results using the

1 Most modern computerized polygraphs do not require periodic field calibration or periodic factory service, though it may be important to check with the manufacturer of each individual instrument. The notion of *functionality check* refers to whether the recording sensors and software are capturing and recording physiological data in the intended ways. Proper functionality is easily observed at the time of an examination when observing and recording stimulus events and changes in sensor activity and can be effectively observed and demonstrated during the administration of an ACQT.

standard card-test and the *known-card-test,* finding no difference in effect sizes for deceptive or truthful outcomes. The known-card-test was also referred to as the *known-solution-test* – a form of *peak-of-tension* test. The known-solution-test differs from the earlier unknown card-test in important ways. The most important difference is that no attempt is made to conceal the examinee's selection when using the known-solution ACQT.

Kirby (1981) also provides insight into the use of the *card-stim* test, also referred to as a *stim-test* or *stimulation-test,* including the emphasis on promoting a perception of the infallibility of the polygraph instrument. The standard-card-test at that time was an *unknown-solution* ACQT for which each examinee was required to select the key question by selecting a number, card or item. The premise was that the examinee would conceal their choice of key question from the examiner. The examiner would then conduct the unknown-solution ACQT and correctly determine the examinee's choice and would then proceed to verbally stimulate the examinee while asserting the infallibility of the test.

This older practice of attempting to assert the infallibility of the polygraph appears to have been premised on an arcane assumption that the effectiveness, or validity, of the polygraph was fundamentally dependent on the examinee's belief that the polygraph was flawless. These assumptions may have contributed to the emergence of field practices wherein every examinee was informed, regardless of the outcome, that they had been correctly detected. Other field practices involved the manipulation of the purportedly unknown-key question such that the examiner either determined the examinee's selection or was fully informed and knowledgeable about the examinee's choice prior to the ACQT. Elliott, Egan & Grubin (2017) provide evidence that it is not necessary for the examinee to believe the complete infallibility of the polygraph for it to be effective.

Reliance on manipulation and deception in the demonstration of polygraph validity is scientifically questionable. Moreover, reliance on manipulation and deception in the ACQT has been described by psychologists as ethically questionable (see Lykken, 1981; Note 2 in Bradley & Janisse, 1981) and potentially problematic in that some examinees may learn of the use of misinformation and manipulation in the ACQT (Ben-Shakhar & Furedy, 1990; Lykken, 1981). One view of type of test represents a form of role-reversal – wherein the examinee is testing the examiner.

Perhaps the most problematic aspect of any attempt to use the ACQT to demonstrate the polygraph's effectiveness to the examinee, is that virtually no well-informed person today believes the polygraph, or any scientific test, to be infallible. Yet studies involving examinees who were knowledgeable or informed about polygraph methodologies (Honts & Reavy, 2009; Honts & Alloway, 2007; Honts & Handler, 2015; Nelson, Handler, Blalock & Hernandez, 2012; Rovner, 1986) have reported effect sizes for test accuracy that are similar to other studies with more naive examinees. This complication to the older intended usage of the ACQT – involving the demonstration or proof of the infallibility of the polygraph – is that highly motivated examinees may be likely to respond with superficial cooperation, and examiners may be at risk for mistaking this for authentic or genuine rapport (though most polygraph examiners would assuredly deny ever making such an error).

Studies on the ACQT and the use of feedback are informative but provide somewhat mixed information. Using a card-test, Ellson, Davis, Saltzman, and Burke (1952) showed that detection of deception was more difficult and reduced on subsequent trials when examinees were informed that deception had been detected in response to the first trial. Later, Davis (1961) hypothesized that guilty subjects might become less psychologically reactive, and therefore less detectable, if they are convinced that their deception is clearly indicated. It is also possible that guilty subjects alter their strategies for concealing their deception after receiving effective feedback.

A subsequent study by Barland and Raskin (1975), using a comparison question test format, showed that the manipulation of feedback, in terms of effective detection of deception, ineffective detection, and no feedback conditions produced an effect for examinee confidence in the polygraph test but failed to produce an effect for the classification of deception or truth-telling. A later study by Horowitz, Kircher and Raskin (1986) who also showed lower rates of deception when examinees were told that their deception had been identified during a number card-test. These findings differ from those of later studies that showed either no effect or desirable effects from the ACQT using a known-solution approach.

An important difference between the unknown-solution ACQT and known-solution ACQT is that the use of feedback with the known-solution ACQT does not attempt to assert the infallibility of the polygraph test, but instead attempts to provide the examinee with feedback as to the effectiveness of the sensors at capturing and recording changes in physiology in response test stimuli. A study by Bradley and Janisse (1981) – in addition to using an electric shock paradigm that showed no effect for the type or intensity of consequences for failing a polygraph test – found high rates of accuracy for the card control test, leading them to suggest that providing the actual results of the card-control test would be an effective approach that is less scientifically and ethically and complicated.

Kircher, Packard, Bell and Bernhardt (2001) studied the effects on the subsequent outcomes of comparison question tests with probable and directed lie questions when subjects were first placed in effective-feedback, ineffective-feedback, and no-feedback conditions using a known-solution ACQT test. Effective feedback in this usage was limited to statements about the observance of a physiological response to the test stimuli. The test structure was a known-solution ACQT format that was previously described by Podlesny and Truslow (1993). However, whereas Podlesny and Truslow instructed the subjects to answer truthfully to all questions, Kircher *et. al.* (2001), instructed the subjects to answer *NO* to all questions including the selected number. Horneman and O'Gorman (1985) previously reported that denial of a selected card led to increased response and detection of the correct number-card compared to affirmative answers or non-answering. Kircher *et al.* reported significant effects for both effective-feedback and no-feedback and generalized the recommendation of Bradley and Janisse (1981) for the use of effective-feedback with the known-solution ACQT.

How to Conduct the ACQT

ACQT Question List

Use of the ACQT begins with the construction of the list of stimulus questions. A commonly used form of the known-solution ACQT involves the use of sequence of numbers. Table 1 shows an example list of questions for a known-solution number test, using the number 4 as the known-key question. The examinee will be instructed to answer *NO* to the known-key question, along with all other ACQT questions. The example in Table 1 includes three buffer questions before the known-key question and three buffer questions after the key-question. Also, the key-question should be clearly indicated in the question sequence.

Table 1. Question list for known-solution ACQT number test.

Question Tag	Type	Question	Answer
X		This practice test is about to begin. Please sit still. Look straight ahead. Listen carefully, and answer "no" to each question. No other talking, and do not move during this practice test.	-
1		Did you write the number 1?	No
2		Did you write the number 2?	No
3		Did you write the number 3?	No
4K		Did you write the number 4?	No
5		Did you write the number 5?	No
6		Did you write the number 6?	No
7		Did you write the number 7?	No
XX	End	This practice test is complete. Please remain still until I release the pressure in the cardio cuff.	-

Table 2 shows another common form of known-solution ACQT, using the examinee's surname as the known-key question. The ACQT surname has been described by field practitioners as a simple and easy known-solution ACQT format for which the salience of the personalized known-key question differs from the other questions. Other known-solutions may also exist, including variants that make use of personal or novel information. Because there is no 'scientific magic-sauce' in the ACQT topic itself, there is no reason to expect any difference in the contribution to polygraph outcomes for different variants of the known solution ACQT.

Table 2. Question list for known-solution surname ACQT.

Question Tag	Type	Question	Answer
X		This practice test is about to begin. Please sit still. Look straight ahead. Listen carefully, and answer "no" to each question. No other talking, and do not move during this practice test.	-
1		Is your surname Johnson?	No
2		Is your surname Jefferson?	No
3		Is your surname Wilson?	No
4K		Is your surname Nelson?	No
5		Is your surname Iverson?	No
6		Is your surname Stevenson?	No
7		Is your surname Mickelson?	No
XX	End	This practice test is complete. Please remain still until I release the pressure in the cardio cuff.	-

Although some examiners may prefer to use a shorter list of ACQT questions, with fewer buffer questions, the use of three buffer questions is recommended because it provides more opportunity for any instruction or admonition needed to improve the examinee's posture or cooperation prior to the presentation of the known-key question. Use of three buffer questions will often ensure that at least one question was presented without the need for in-test instruction prior to the known-key question.

Examiners who use less than three buffer questions before or after the known key question will have more limited insight as to how the examinee will cooperate and respond during testing. In the same way that shortening the question interval can reduce the length of the ACQT, use of fewer ACQT questions may incorrectly instruct the examinee as to the expected length or duration of the question sequence during the data-collection phase of the polygraph test.

Many examinees are unaccustomed to polygraph testing, and may produce what is termed artifacted, or unstable data at the onset of testing due to unsatisfactory posture, cooperation (i.e., excessive movement) or distraction. These examinees may benefit from additional in-test instruction. It is common for some examinees to move slightly upon the X announcement of test onset, and some examiners may instruct the examinee at that time.

Other examinees may move upon answering the first question. This is also an opportunity to provide instruction to improve the examinee's behavior and cooperation during test. Another possible problem is that some examinees may become distracted during the silent periods between questions. This is also an opportunity for an examiner to provide information that helps motivated and truthful examinees to cooperate successfully.

Use of three buffer questions prior to and after the known-key question will ensure that the ACQT is closer in length to the actual polygraph test question sequence and will better orient the examinee as to what to expect during testing. This provides an opportunity for examinees to practice sitting patiently throughout the testing process and provides an examiner with more opportunity to observe the examinee's posture and cooperation during silent periods in between the test stimuli. Use of fewer buffer questions provides less opportunity for practice and observation. There is little value in the use of short-cut procedures during the ACQT.

Introduction and Review of the ACQT Questions

Introduction of the ACQT questions begins when the examiner informs the examinee that a practice test will be completed, and then requests the examinee to write a number (i.e. *"4")* in large print in the middle of a large circle which the examiner has drawn on a sheet of blank paper. The examiner will use this activity to orient or sensitize the examinee's attention to the number they chose, in this example *4*, and this can sometimes be done by circling the number *4* several times, or by asking the examinee to print boldly or over-print the number one or more times. The examiner will then, in the presence of the examinee, write the numbers *1, 2* and *3* before the number *4*, and then proceed to write the numbers *5, 6* and *7* after the number *4*. The examiner should explain to the examinee that it is obvious that the examiner has written the numbers *1, 2*, and *3*, and the numbers *5, 6*, and *7*, while the examinee is the person who has written the number *4*.

When using the surname test, having the examinee spell the surname, letter by letter, is good practice, as it can help to avoid spelling and documentation errors and can also serve to orient or sensitize the examinee's attention to the known-key question. Other variations on the known-solution ACQT may employ other methods to orient and sensitize the examinee to the known-key question, including simple math questions. Some examiners may choose to display a paper with the list of question items in front of the examinee during the ACQT. Others may request the examinee to fold the paper and sit on it during the ACQT. These activities are non-critical and are intended only to contribute to the examinee's heightened attention and awareness of the ACQT questions.

The examiner will explain to the examinee that during the test, he/she will be asked if they wrote the number 1, 2 and 3 (i.e., *Did you write the number 1? Did you write the number 2? Did you write the number 3?*)[2]. The examiner should read each question to the examinee and allow the examinee to answer. The examinee will normally answer *NO* to each question, and the examiner will advise that these answers are known to be truthful because the examiner, not the examinee, has written those numbers. The examiner may explain further that these questions provide an opportunity for the examiner to observe normally expected physiological responses.

The examiner will then present the known-key question, (*Did you write the number 4?),* along with a clear instruction that the examinee is to answer NO to this question, in the same manner as he/she has answered *NO* to each of the previous questions. The examiner should emphasize that it is already known that the examinee did in fact write the number 4 and that the required *NO* answer is incorrect. The examiner will further advise that the purpose of this is simply to observe how the examinee responds physiologically[3]. Some examinees may inquire to clarify that they are being instructed to lie, and it is acceptable to refer to the *NO* answer as a lie if the examinee does so.

Some examinees will know answering *NO* is not actually lying when one is instructed to answer *NO*. It is simply an instruction – a procedure – but not a lie. To avoid potential complications, it may be useful for the examiner to avoid referring to the requested *NO* answer as a *lie* unless the examinee spontaneously refers to it as a lie. Examiners who maneuver the examinee into referring to the requested answer as a lie may be at risk for mistaking superficial cooperation for rapport.

Other examinees may attempt to object to the idea of telling any lies during any part of the polygraph process or at any other times – asserting this to be inconsistent with personal values or religious beliefs. In these cases, it is often best to simply advise the examinee that answering *NO* in this context is not a lie *per se* but is merely an instruction and a procedure, and that failure to follow the instructions and procedures may result in not passing the polygraph test.

The examiner will complete the review of the ACQT questions by reading each of remaining questions and allowing the examinee to answer. The examiner should advise the examinee once again that it is already known that these answers are truthful or correct because the examiner has written the numbers 5, 6 and 7. The examiner should then advise the examinee that the purpose of the practice test is to observe the response when the examinee answers *NO* to the questions about number 4.

Introduction of the ACQT with the surname or other topics is like the use of the ACQT with numbers. The examinee is instructed to answer known-key question incorrectly with a *NO* answer. Also, it may be necessary to alter the buffer questions to exclude the known-key item when using a variation to the number-test.

2 Another form of these question has also been used, in this manner: *"Regarding the number you wrote, was it the number 1? "Was it the number 2? "Was it the number 3?"* Although semantically identical this wording has less similarity with common language usage. A principle of polygraph questions formulation is to use of comfortable and common language whenever possible to avoid reactions that may result from novelty or confusion from uncommon language usage.

3 It is not advisable to state that the purpose of this exercise is to observe what it looks like or what the examinee's body does when lying – because this statement would be obviously factually incorrect. The basic physiological patterns – phasic change and return to tonicity – is similar for many types of responses. Also, the act of lying to a matter under investigation is thought to be a distinct phenomenon from answering NO incorrectly when instructed to do so. For these reasons, it is advisable to limit this statement to one that is simple and factually correct: the purpose of the activity is to observe the physiological response.

Best practice is for the examiner to refrain from taking any shortcuts in the documentation and preparation of the list of polygraph test questions. This means that examiners should always type or write each of the questions completely, without shorthand, using correct spelling, punctuation, and use of capitalization.

After introducing and ascertaining the examinee's answer to the first three surname buffer questions, the examiner should alert the examinee that the next question will be the correct surname question but that it is a requirement to answer *NO*. The examiner should then ascertain through practice that the examinee can answer *NO* to the known-key question. The examiner can then introduce and ascertain the examinee's answer to the remaining buffer questions. The examiner should then advise the examinee that the purpose of the practice test is to observe the response when the examinee answers *NO* to the questions about his or her surname.

ACQT Question Interval

It is recommended that the ACQT is conducted with the same 25 second question pace interval as the data-collection phase of the polygraph test. While it may be tempting for some field practitioners to shorten the question interval for the ACQT – with the view that it is un-important or less important – use of a shorter interval may be problematic for several reasons. For example, use of a shorter question interval during the ACQT may deprive the examiner of an opportunity to fully observe the examinee's normal posture and cooperation after each test question and before the next stimulus is presented. Also, shorter question intervals may provide insufficient time for the physiological data to return to the tonic level before each subsequent test stimuli. A shortened question interval may lead some examinees to expect a similarly short interval during the polygraph data collection phase, leading to unintended cognitive activity or other reactions for examinees who notice what appears to them be an unusually long wait – though it is the normal intended interval – in between questions during the actual polygraph test.

Attachment of Polygraph Recording Sensors

The polygraph recording sensors should be attached to the examinee following the review of the ACQT questions. The examiner should briefly remind the examinee about the purpose or function of each sensor, in addition to advising the examinee about the importance of remaining still during testing. Many polygraph examinees will have an under-developed understanding of how to sit still during testing. It may be helpful for the examiner to clearly advise the examinee about how to remain still including the importance of keeping one's feet flat on the floor, allowing the chair to support one's posture, and keeping one's arms on the arm-rest or desk – in addition to the importance of looking straight ahead during testing. The examiner should take notice of those examinees who may benefit from an additional support to stabilize their feet during testing. Also, it is ideal if the examinee can use the back of the chair to further support and stabilize his or her head and posture during testing.

Examinee's should be advised to keep their eyes open during testing. This will help the examinee to avoid falling asleep during testing and may also help to avoid problems from increased attention to other physical sensations when one's eyes are closed, or problems related to past issues of trauma for some examinees. Using a visual focal point or visual reference point may help some examinees to refrain from looking around or moving their head during testing.

Although it does not affect the scientific validity of a test result, polygraph recording sensors are normally attached in a consistent sequence, beginning with the lower and upper respiration sensors. The traditional procedure is for the examiner to provide simple instruction to reposition the examinee while attaching the respiration sensors. Some examinees may experience less social discomfort if they are instructed as how to attach the sensors to themselves. The cardio cuff is normally attached to the examinee after the respiration sensors.

There is no empirical evidence to support a requirement that the cardio sensor or other sensors must be attached to the right or left side of the examinee's body. Subsequently, there is no evidence regarding the direction of the cardio tubing. What is most important is that the cardio sensor and tubing must be located in a manner in which it will neither be disturbed by nor distract the examinee during the testing process. Electrodermal sensors can be attached to the examinee's right or left side as is most convenient for the testing location. There is no scientific evidence to suggest that the choice of right or left side has any effect on polygraph outcomes. Similarly, a fingertip vasomotor sensor can be attached to either the right or left hand. There is no basis of evidence to support requirements for the separation of recording sensors to different sides. Rules or constraints on these matters would increase examiner venerability to criticism with no known effect on scientific validity of the test data and test result. Most importantly, highly skilled examiners will be able to attach the polygraph sensors without placing the examinees in socially awkward and/or physically uncomfortable positions.

ACQT Recording

After all recording sensors are attached, the examiner will make any necessary adjustments and then initiate the ACQT recording. The examiner should inform the examinee that the recording has begun and will begin to advise the examinee of the need to sit still. Some examiners will inflate and stabilize the cardio cuff sensor prior to initiating the recording. This will generally result in less visual distortion at the onset of recording. Other examiners may prefer to start the recording before inflating the cardio cuff. Here the advantage is that the process of inflating and stabilizing the cardio cuff is permanently included in the recorded data and subsequently available for visual review.

Additional adjustments may be made to the data following the onset of recording. The ACQT itself begins with the announcement of test onset (X) and will proceed through the ACQT question sequence until the announcement of test completion (XX). The recording is stopped following the completion of the ACQT data collection, after which the data can be dressed or adjusted for optimal visual display and then saved to the computer storage device.

If any of the recording sensors are not functioning normally, or if the examiner is unable to stabilize the data to a satisfactory and usable degree, the examiner should terminate the recording and then correct the problem. After the problems are corrected, the recording can be started again and the ACQT completed. If an examinee's physiological data is observed to be of insufficient quality or stability, an examiner may elect to forgo polygraph testing or may wish to consult with the referring agent before deciding how to proceed.

Examinees who move excessively or who experience distraction during the ACQT can be given simple instructions. For example: "it is important that you stay still during the test," or "it is important that you concentrate and do not get distracted during the test." Similarly, problems involving disruptive deep breathing can be calmly addressed by advising "it is important that you do not move your upper body during this practice test." Instructions of this type should not be repeated more than two times.

Effective in-test instruction and skillful management of observed problems during the ACQT procedure may give the examiner an opportunity to observe whether the examinee is capable of and willing to cooperate during testing. An examiner may choose to abort and restart the ACQT if it necessary to provide an examinee with additional information or instruction in response to observed problems with attention, posture or cooperation. It may also be acceptable for an examiner to continue the ACQT to completion even after continued problems are observed after advising the examinee. Under some conditions an examiner may elect to repeat the ACQT.

If necessary, the ACQT may be aborted to address and correct any observed problems with the functioning of the recording sensors. It is unproductive to complete the ACQT if the sensors are not functioning as intended. The ACQT should be restarted after the problem is corrected. Correct functionality will be observed in the form of normal physiological activity in respiration, cardio, electrodermal, vasomotor and activity sensors, along with observable changes in activity in response to the ACQT stimuli.

Known-Key Question

The known-solution ACQT is not dependent on psychological or situational manipulation, and for this reason can be viewed as less ethically and scientifically controversial. The known-solution ACQT permits more potential for standardization than the unknown-solution test, including the potential for the use of the same known-key question, located in the same position in the ACQT question sequence, for each examinee. Examiners who are tempted to add variation to the known-key item or ACQT question sequence are cautioned against adding variation only to relieve occasional professional boredom and are advised to embrace the value of a consistency in the applied use of the ACQT.

Consistency in the administration of the ACQT will help to avoid errors. Consistent administration of the ACQT will allow an examiner to gain more insight about individual differences in behavior and response at recording onset, attention during silent periods, response to in-test instructions, and response to test stimuli both before and following the known-key question. Although this information is unquantified and not subject to objective analysis, it is a potentially rich source of information that can assist an examiner to work effectively with each individual examinee.

Some examinees may answer incorrectly in response to the known-key question. If the examinee does not answer the known-key question as instructed, the examiner may terminate the ACQT and advise the examinee of the error and need to answer as instructed. It is possible that the examinee forgot instructions, though it is also possible that some examinees may choose not to cooperate. In either case the examiner may be able to rectify the problem by re-instructing the examinee and re-starting the ACQT. Some examiners may elect to provide the examinee with in-test instruction regarding the requested answer and then repeat the known-key question, either before presenting the remaining buffer questions of before the XX announcement of test completion. Because this is simply an acquaintance test, there is no known reason why this should be considered unacceptable.

The known-key question will normally produce an observable change in physiological activity. If no response is observed, the examiner may elect to repeat the known-key question at the end of the ACQT sequence, before the announcement of test completion. Though repetition is usually un-necessary, it may be useful to repeat the ACQT under some circumstances, such as when the examinee does not cooperate in a satisfactory manner or when the data are unsatisfactory and can be improved with some adjustment to the sensor or instrument. Some examinee's physiological responses may be unusable or un-interpretable. In these cases, a decision to proceed with testing under the prior knowledge that the data are unusable or uninterpretable would not be without some ethical complication.

ACQT Feedback

Common practice is to provide the examinee with feedback after the completion of the ACQT. Use of the known-solution ACQT – which cannot be confused with any form of parlor-trick – permits a standardized approach to ACQT feedback. It is reasonable to inform the examinee that he or she has shown a response to the known-key question. It is also reasonable to advise the examinee that he or she is likely to show a reaction in response to deception during the actual polygraph test, and similarly reasonable to advise that there should be no difficulty if he or she is telling the truth about the matter under investigation. An example is shown below:

> *You showed a reaction to that question about (the number 4 / your name / other ACQT topic). If you lie during the polygraph you are likely to show a reaction. On the other hand, there should be no great difficulty determining if you are telling the truth.*

These statements make no claims about the detection of deception or the magnitude of physiological response. And, empirical data supports these statements for normal functioning examinees. Detection of deception may not be a realistic endeavor with a single iteration of a question sequence, and the use of a known-solution ACQT creates a context in which no actual deception has occurred.

The purpose of the known-solution ACQT is to allow the examinee to become accustomed to the sensors, testing procedure, need for cooperation, and for the examiner to ascertain that the instrument and sensors are functioning as intended. For this reason, it is neither necessary nor advisable to tell the examinee that any deception has been detected. Nor it is advisable to tell the examinee anything about the difference or size of the observed reactions.

It is not advisable to show the ACQT charts to the examinee, because providing this information may contribute to a change or increase in the way the examinee attends to his or her perceptions and awareness of physiological activity and responses during testing. Examinees who desire to engage in countermeasures during testing may attempt to misuse the information gained if an examiner shows the test data. This view contrasts with older practices that sometimes involved showing the ACQT data to the examinee[4].

Summary

The ACQT has been used by polygraph examiners since the early history of the polygraph profession, though its use has changed somewhat over the years. Along with subtle but important changes in ACQT field practices, some changes have occurred in the terminology used to refer to the ACQT. For example, some early practices involved the use of the ACQT between the first and second test chart and referred to the ACQT as a *stim test* or *stimulation test*. The term *stimulation* may have been thought by some to be problematic, and the profession has gravitated away from that term in the form of more comfortable and general words like *acquaintance test* or simply *practice test*. The method described herein is highly standardized, and applicable to a wide range of examinees and testing contexts. It does not depend on parlor-tricks, manipulation or misinformation, and is consistent with published scientific evidence on the beneficial effects of the ACQT on polygraph outcomes.

The known solution ACQT is the only form of acquaintance test described in available publications on polygraph field standards of the U.S. federal government (Department of Defense, 2006). Considering the available scientific evidence for the known solution ACQT there is little argument that it is a valuable part of the test, and little argument for the use of any unstudied or experimental form of ACQT in lieu of the evidence-based known-solution test. It is consistent with published evidence and applicable to a wide variety of polygraph screening and diagnostic contexts, including criminal investigations, public safety applicant screening, employee screening, security screening, and post-conviction testing. The known-solution ACQT is highly standardized and does not rely on psychological or situational manipulation, or misinformation.

There is much to learn about an examinee from the careful and competent use of a known-solution ACQT. In addition to ensuring that the instrument and recording sensors are functioning as intended, it is an opportunity to observe the examinee's posture and cooperation during testing, and potentially rectify any problems before the actual polygraph examination. Skillful use of the known-solution ACQT has been shown to increase the effectiveness of the polygraph test. The actual reason for this effect appears to have little to do proving or demonstrating the effectiveness or infallibility of the polygraph test and may have

4 The most concerning of all old-school manipulative ACQT practices that the authors are aware of involves the increase of test sensitivity adjustment at the time of the key-question, resulting in a response that would always be visually impressive though unrepresentative of the actual change in physiological activity. This practice is not possible with contemporary computerized polygraph instruments.

more to do with ensuring that the instrument and recording sensors are functioning properly and that the examinee has had an opportunity to practice cooperating with behavioral instructions during the polygraph test.

References

American Polygraph Association (2018). *APA Standards of Practice (Effective March 24, 2018. [Electronic version]* Retrieved August 17, 2018, from http://www.polygraph.org.

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

Ben-Shakhar, G. & Furedy, J. J. (1990). Theories and applications in the detection of deception: A psychophysiological and international perspective. Springer-Verlag.

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

Department of Defense (2006). *Federal Psychophysiological Detection of Deception Examiner Handbook.* Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007. Reprinted in *Polygraph*, 40(1), 2-66.

Elliott, E., Egan, V. & Grubin, D. (2017). A Not So Bogus Pipeline: A Study of the Bogus Pipeline Effect and Its Implications for Polygraph Testing. Polygraph,46(1), 1-9.

Ellson, D. G., Davis, R. C., Saltzman, J. A. & Burke, C. J. (1952). A report of research on detection of deception. (Contract N. N6 ONR 18011, Office of Naval Research). Bloomington: University of Indiana.

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 [abstract].

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. & Reavy, R. (2015). The comparison question polygraph test: A contrast of methods and scoring. *Physiology and Behavior, 143,* 15-26.

Honts, C. R., Handler, M., Shaw, P., & Gougler, M. (2015). The Vasomotor Response in the Comparison Question Test. *Polygraph, 41(1),* 62-78.

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

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.

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

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.

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.

Reid, J. E. & Inbau, F. E. (1966). *Truth and deception: The polygraph ("lie detector") technique.* Baltimore: Williams & Wilkins.

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

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

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.

On Standard Methods of Forensic Psychophysiological Examination Using a Polygraph[1]

Komissarova Yaroslava V.

Assistant Professor of the Department of Criminalistics of the Kutafin Moscow State Law University (MSAL)

Editor-in-Chief of the Forensic Expert Federal Scientific and Practical Journal

Candidate of Legal Sciences

Assistant Professor

a5143836@yandex.ru

Editorial note: This letter to the editor is in response to the manuscript Kalafati & Krapohl (2018). It is published without edit and with its publication the journal has allowed all sides to have a voice, and that the matter is now closed.

Developments and recommendations of individual researchers may in some cases extend beyond the state of the art and be contrary to the current laws. It is quite appropriate for theoreticians, as discussions are beneficial for science. However, it is unacceptable for investigative practices and court proceedings. To be recognized as standard, forensic methods must be tested and implemented.

Keywords: *polygraph examination, psychophysiological problem, expert methods, procedural legislation.*

Psychophysiological examination using a polygraph (hereinafter referred to as PPE) for the purpose of obtaining an opinion of a polygraph examiner (including PPE under Article 80 of the Criminal Procedure Code of the Russian Federation), polygraph interrogation, or PGI (including a type of an operational search activity "interrogation" tool, which may be conducted using equipment without detriment to people's life or health under Article 6 of Federal Law No. 144-ФЗ of August 12, 1995, on Operational Search Activities), forensic psychophysiological examination using a polygraph (hereinafter referred to as FPPE) are different legal forms of simultaneous applied expertise in psychology, physiology, and forensic science[2].

While the practice of using a polygraph to fight crime is almost 100 years old, in Russia, the question of using data provided by a polygraph examiner as evidence in criminal cases remains open. This was confirmed by the participants of round-table discussion "Problems of Polygraph Use in Investigation and Prevention of Corruption-Related Offenses" held in January 2015 by the Department of Criminalistics of the Law Faculty of the Lomonosov MSU, the Department of Criminal Procedure and Criminalistics of the SPbSU, the Department of Criminalistics of the Kutafin Moscow State Law University (MSAL) in association with Non-Profit Partnership "The National Board of Polygraphologists".

At the opening, the problem was referred to by:

- *I.V. Aleksandrov*, Head of the Department of Criminalistics of the Law Faculty the Lomonosov MSU, Doctor of Law Professor;

- *E.P. Ischenko*, Head of the Department of Criminalistics of the Kutafin Moscow State Law University (MSAL), Doctor of Law, Professor;

- *N.A. Sidorova*, Deputy Head of the Department of Criminal Procedure and Criminalistics of the SPbSU, Candidate of Legal Sciences, Assistant Professor.

- *Svetoslav Zanev*, PhD, Chairman of the Supervisory Board of the Bulgarian Polygraph Association, Director of the International Association of Polygraph Examiners, spoke about successful experience of using a polygraph for psychological examinations.

The following speakers elaborated on the topic of the round-table discussion:

- *Yu.M. Drobyazka*, Director of Non-Profit Partnership "The National Board of Polygraphologists";

- *S.Yu. Aleskovsky*, Professor at the Academy of Economics and Law, President of the Eurasian Polygraph Association (The Republic of Kazakhstan, Almaty);

- *S.A. Aksenov*, Deputy Head of the Classified Information and Material Security Division of the Department of Competition Policy (the Moscow City Government);

- *Ya.V. Komissarova*, Assistant Professor of the Department of Criminalistics of the Kutafin Moscow State Law University (MSAL);

- *V.P. Taskaev*, Deputy of the State Duma of the Federal Assembly of the Russian Federation;

- *V.D. Pristanskov*, Assistant Professor of the Department of Criminal Procedure and Criminalistics of the Law Faculty of the SPbSU.

The participants of the discussion draw attention to a brochure named "Standard Methods of Forensic Psychophysiological Examination Using a Polygraph"[3] (hereinafter referred to as the "Methods") printed in the publishing office of the Moscow Academy of the Investigative Committee of the Russian Federation in Autumn 2014. Neither the authors of the Methods, Yu.I. Kholodny and Yu.K. Orlov, nor the reviewers are officials of the Russian Investigative Committee of the Academy of the Investigative Committee. However, the abstracts state that the Methods are intended for students of postgraduate courses of the Academy of the Investigative Committee and for employees of investigative departments of the Russian Investigative Committee.

During the round-table discussion, it was suggested to hand over the Methods to the members of the Expert Council of Non-Profit Partnership "The National Board of Polygraphologists" for discussion, which resulted in the following scientific and advisory opinion prepared by the author.

1. The statement that the Methods "are a summary of scientific and applied scientific researches and practices of application of a polygraph in operational search and investigative practices accumulated by Russian scientists and experts from 1970 to 2014" (page 5) is not consistent with the reality.

The contents of the publication reflect its authors' personal opinion. In fact, one of them, Yu.I. Kholodny, although familiar with polygraph use, has never conducted polygraph examinations. The other author, Yu.K. Orlov, is an expert in criminal procedures and forensic examination theory rather than a polygraph examiner.

The authors of the Methods use terms, concepts and classifications which are either not present or are deemed obsolete in Russian and international publications known to the reviewers. For example, on page 11, in paragraph 2.3, they propose an original classification of questions, wherein "comparison questions" are referred to as "control (comparative) questions". On page 13, in paragraph 2.7, the authors describe "control tests" which ensure "control of the method correctness" of polygraph test administration. Such test does not exist. Accordingly, the statement that "the number of control tests must be at least 40% of the relevant tests" (para. 3.7.5, page 21) is meaningless.

The publication does not provide a polygraph chart analysis algorithm. It does not specify modern methods for examiners to quantify the probability of accuracy of the conclusions based on the results of an expert assessment of the test results (assess the probability of errors in the conclusions).

The list of references contains only 14 names. These include three works on forensic examination theory and practice in general, the other 11 are written by the Methods authors (one of them or both), 8 of which are small articles.

2. According to the authors of the Methods, a polygraph is a "medicobiological class" device (page 8, paragraph 1.7). There is no such class (classification of devices). <...>

3. Provisions of the Methods (pages 16-30) regarding statement of the problems to be resolved by the polygraph examiner and expert conclusions are not scientifically based, as they are beyond the state of the art.

Obtaining of forensically significant information always involves working with a physical media. In case of immaterial traces, forensic scientists inevitably face the so-called psychophysiological problem. It involves answering the question of correlation between mental and physiological (neurobiological) processes. Currently, the psychophysiological problem is far from being solved even hypothetically; psychology offers several approaches to its resolution[4].

Yu.I. Kholodny attempted to gain an insight into PGI using his own targeted memory testing theory. According to Yu.I. Kholodny, the results of a polygraph test allow to make a conclusion about availability of specific information in a person's memory[5]. He reduces the very complex, underexplored process of formation of memory traces to purely physiological processes (also understudied).

Once, a well-known Soviet psychologist, A.N. Leontyev, wrote that it was impossible to reduce psychological laws down to brain function laws[6]. Now, several decades later, his postulate remains true. Correlation between the brain and psyche is of a system rather than linear nature: "it is impossible to map any psychic manifestation directly to some physiological processes of the brain in principle"[7].

Yu.I. Kholodny's opinion faced reasoned criticism from experienced polygraph examiners[8]. While, admittedly, higher mental functions (such as memory, emotions, thinking process, etc.) play some part in the formation and development of people's psychophysiological responses to external and internal stimuli, it has to be said that memory processes as such are not directly related to the autonomic nervous system functioning mechanisms the external manifestations of which in the form of physiological changes (responses) are recorded with a polygraph. "There is a huge distance between them, and within this distance, there are many things, such as emotions, cognitive processes and various purely physiological phenomena, preventing the information stored in a person's memory from being constantly and unambiguously reflected in external physiological manifestations that would allow us to come to any accurate conclusions regarding the availability or presence of such information"[9].

By actualizing images stored in the test subject's memory (including by presentation of stimuli, selected and systematized in a certain order) polygraph examiners evaluate the significance, stability, ratio of responses to the test questions. By using various systems of evaluation of recorded data, they can determine which stimuli are relevant to the test subject. Depending on the type of stimuli and the technique used for polygraph testing, polygraph examiners can conclude whether any responses were detected during examination, that indicate that the test subject bears knowledge about an event (or details thereof).

Thus, based on the analysis of the detected response, polygraph examiners provide their (expert) version of the test subject's knowledge about the event. Having expertise, they can also provide an opinion on how the test subject may have had obtained this information about the event (the probability of it being obtained at the time of the event). However, polygraph testing results do not reveal what information exactly is stored in a person's memory because memory mechanisms are understudied at present.

4. Taking into account the principles of Russian criminal proceedings, in modern conditions the use of so-called control questions of a "guilt complex" is unacceptable (page 38, para. 23 (f)).

"Guilt complex" questions are questions regarding a hypothetical (fictional) event topically relevant to the crime being investigated. In the mid 20th century, American experts suggested that using such questions may cause detection of an inadequate response (due to a number of factors) to the polygraph testing procedure in the test subject (the mere fact of being suspected or accused of committing illegal actions)[10].

Forensic tactics enable practical application of techniques based on delusion of the accused as to what the investigator truly knows about the actual circumstances. However, these techniques have nothing to do with the lie on which the "guilt complex" questions are built. Due to the lack of insight into all memory mechanisms and principles of professional conduct of the expert, such manipulation and suggestion methods are unacceptable.

5. Some recommendations of the authors of the Methods contradict the provisions of the current Russian laws. <…>

PS. 1 (published with the article in 2015). In March 2015, the scientific and advisory opinion signed by the author of the article and the letter about scientific inadequacy of the Methods signed by V. N. Fedorenko, a famous polygraph expert, received by Non-Profit Partnership "The National Board of Polygraphologists" were submitted by V.P. Taskayev, a Deputy of the State Duma of the Federal Assembly of the Russian Federation, to the Chairman of the Investigative Committee of the Russian Federation with the following inquiry:

"Have the Standard Methods of Forensic Psychophysiological Examination Using a Polygraph developed by Yu.I. Kholodny and Yu.K. Orlov been tested or not (if yes, specify the place and duration)?"

This question is extremely important as forensic theoreticians and practitioners rightly state: "To be recognized as standard, forensic methods must be tested and implemented"[11]. Even the co-author of the Methods, Yu.K. Orlov, only recently wrote that the reliability of an expert method must be primarily evaluated on formal grounds: "the trustworthiness of the developer institution, testing time and organization, the authority that approved and recommended it for use"[12], etc.

According to the response, the brochure by Yu.I. Kholodny and Yu.K. Orlov entitled "Standard Methods of Forensic Psychophysiological Examination Using a Polygraph" printed in the publishing office of the Academy of the Investigative Committee of the Russian Federation reflects its authors' opinions regarding concepts, theories of and procedures for FPPE:

"The Academy of the Investigative Committee of the Russian Federation provides no evaluation of the method, its advantages or deficiencies compared to other methods, as this falls within the competence of expert institutions and organizations". Information and reference publications of the Academy of the Investigative Committee of the Russian Federation are not distributed among other departments.

PS. 2 (to the published translation). The Methods of Yu.I. Kholodny and Yu.K. Orlov (deceased in 2016) have not been tested by the Investigative Committee of the Russian Federation. Currently, tests and examinations involving the use of a polygraph are conducted by the Investigative Committee, the 111 Main State Center for Forensic Medical and Criminalistics Expert Examination" of the Ministry of Defense of the Russian Federation, forensic divisions of internal affairs agencies of Russia, a number of non-governmental expert institutions using: "Generic Expert Methods of Psychophysiological Examination Using a Polygraph" (2005) and the Standard Requirements to the Procedure for Psychophysiological Examinations Using a Polygraph (2008). Scientific research "Procedure for Commissioning and Conducting Psychophysiological Examinations and Tests Using a Polygraph in the agencies of the Ministry of Internal Affairs of the Russian Federation" (Komissarova Ya. V., a co-author) is being completed. It will provide new guidelines for conducting polygraph examinations by forensic divisions of internal affairs agencies of the Russian Federation.

References

1. Alekseev L.G. Psychophysiology of Lie Detection. Methods / L.G. Alekseev – Moscow, 2011.

2. Varlamov V.A. Computer-Assisted Lie Detection / V.A. Varlamov, G.V. Varlamov. – Moscow, 2010.

3. Instrumental “Lie Detection” : Academic Course / S.I. Ogloblin, A.Yu. Molchanov. – Yaroslavl, 2004.

4. Komissarova Ya.V. Tasks, Object and Scope of Forensic Psychophysiological Examination Using a Polygraph / Ya.V. Komissarova // Forensic Scientist’s Library. Scientific Journal. 2012. № 3 (4). S. 251-275.

5. Komissarova Ya.V. Professional Activities of an Expert in Criminal Proceedings: Theory and Practice / Ya.V. Komissarova. – Moscow, 2014.

6. Komissarova Ya.V. Polygraph in Russia and the USA: Application Problems / Ya. V. Komissarova, N.I. Myagkikh, A.B. Pelenitsyn. – Moscow, 2012.

7. Leontyev A.N. Selected Works on Psychology: in 2 vol. Vol.II. / A.N. Leontyev. – Moscow, 1983.

8. Molchanov A.Yu. General Theory of Polygraph Tests / A.Yu. Molchanov, A.Yu. Babikov. — Yaroslavl, 2012.

9. Obukhov A.N. Theoretical and Procedural Fundamentals of Polygraph Use / A.N. Obukhov, I.P. Obukhova — 3rd Edition – Domodedovo, 2012.

10. General Psychology: in 7 vol. : Textbook for Students of Higher Educational Institutions / edited by B.S. Bratusya. Vol. 3. Memory / V.V. Nurkova. – Moscow, 2006

11. Orlov Yu.K. Forensic Examination as an Instrument of Proof in Criminal Proceedings : scientific publication / Yu.K. Orlov. – Moscow, 2005.

12. Pelenitsyn A.B. So What Does the Polygraph Detect? / A.B. Pelenitsyn, A.P. Soshnikov, O.V. Zhbankova // Bulletin of Criminalistics. 2011. Issue 2 (38). Moscow, 2011. S. 7-18.

13. Polygraph in Russia: 1993 to 2008 : Retrospective Collected Works / compiled by Yu.I. Kholodny. – Moscow, 2008.

14. Popovichev S.V. Easy to Lie Hard. Instrumental Lie Detection: from Ideology to Technology / S.V. Popovichev. – Moscow, 2011.

15. Rossinskaya E.R. A Comment on the Federal Law on Forensic Activities in the Russian Federation / E.R. Rossinskaya. – Moscow, 2002.

16. Semenov V.V. Legal, Tactical and Procedural Aspects of Using a Polygraph in Criminal Proceedings : Textbook / V.V. Semenov, L.N. Ivanov. – Moscow, 2008.

17. Soshnikov A.P. Polygraph in Crime Investigation Practices. Methodological Guidelines / A.P. Soshnikov, Ya.V. Komissarova, A.B. Pelenitsyn, V.N. Fedorenko. – Moscow, 2008.

18. Forensic Examination Theory : Textbook / E.R. Rossinskaya, E.I. Galyashina, A.M. Zinin ; edited by E.R. Rossinskaya. – Moscow, 2009.

19. Kholodny Yu.I. Standard Methods of Forensic Psychophysiological Examination Using a Polygraph / Yu.I. Kholodny, Yu.K. Orlov. – Moscow, 2014.

20. Chekhovskikh M.I. Psychology : Textbook / M.I. Chekhovskikh. – Moscow, 2003.

21. Ekman P. Psychology of Lies / P. Ekman. – Saint Petersburg, 2003.

About the author:

Komissarova Yaroslava Vladimirovna, PhD, docent, Associate Professor of the Criminalistics Department in the Kutafin Moscow State Law University (MSAL), Editor-in-chief of the Federal science-practice journal "Forensics analyst" (Moscow, Russia). Yaroslava Komissarova can be reached at a5143836@yandex.ru

---

(Endnotes)

1 The article was published in a peer-reviewed journal: Legal Psychology. 2015. № 3. S. 7-11. Translated with some abridgments.

2 According to some sources, the area of scientific and applied research related to theoretical justification of the effectiveness, as well as actual application of a polygraph is called polygraphology (see Komissarova Ya. V., Myagkikh N.I., Pelenitsyn A.B. Polygraph in Russia and the USA: Application Problems. Moscow, 2012. S. 57).

3 Kholodny Yu.I., Orlov Yu.K. Standard Methods of Forensic Psychophysiological Examination Using a Polygraph. Moscow, 2014

4 General Psychology: in 7 vol. : Textbook for Students of Higher Educational Institutions / edited by B.S. Bratusya. Vol. 3. Memory / V.V. Nurkova. Moscow, 2006. S. 272-275 (Note: Equivalent of the Russian publication: Alan D. Baddeley, Michael W. Eysenck, Michael C. Anderson. Memory. Psychology Press, 2009. 451 s.)

5 Polygraph in Russia: 1993 to 2008 : Retrospective Collected Works / compiled by Yu.I. Kholodny. Moscow, 2008. S. 45-47, 87-90

6 Leontyev A.N. Selected Works on Psychology: in 2 vol. Vol.II. Moscow, 1983. S. 123

7 Chekhovskikh M.I. Psychology : Textbook. Moscow, 2003. S. 8-9

8 There is a number of theories seeking to describe the nature of reactions detected by polygraph tests. In the early XXI century, upon request of the United States Department of Energy, as instructed by the US Government, a Committee to Review the Scientific Evidence on the Polygraph was established under the auspices of the National Academy of Sciences. Upon fundamental analysis of almost all PPE aspects that was performed by experts for 19 months, American scientists concluded that “theoretical justification of the use of polygraphs is rather poor”, and various theories can be worthwhile in different situations. (See: Komissarova Ya. V., Myagkikh N.I., Pelenitsyn A.B., ref. above, S. 50).

9 Pelenitsyn A.B., Soshnikov A.P., Zhbankova O.V. So What Does the Polygraph Detect? // Bulletin of Criminalistics. Issue 2 (38). Moscow, 2011. S. 8-9.

10 Instrumental “Lie Detection” : Academic Course / S.I. Ogloblin, A.Yu. Molchanov. Yaroslavl, 2004. S. 267.

11 Forensic Examination Theory : Textbook / E.R. Rossinskaya, E.I. Galyashina, A.M. Zinin ; edited by E.R. Rossinskaya, Moscow, 2009. S. 132

12 See: Orlov Yu.K. Forensic Examination as an Instrument of Proof in Criminal Proceedings : scientific publication. Moscow, 2005. S. 143–144

Polygraph Questions and Questionable Questions:

Sexual Thoughts and Fantasies

in Post-Conviction Treatment and Supervision

Raymond Nelson[1]

Abstract

Central to the effectiveness of the polygraph, or any test, are the test stimuli. This paper argues that the validity of the polygraph test has been established as a test of behavior but not as a test for thoughts and fantasies independent of behavior. Potential problems in polygraph target selection and question formulation are discussed in the post-conviction sex offender testing context, with attention to the capabilities and limitations of the polygraph test. The polygraph does not measure lies *per se*, is not a test of *mens rea*, cannot read minds, and is not a test of intent or intention. Caution is urged around the unscientific use of the polygraph in attempt to test for unreported sexual thoughts and fantasies that are not expressed in behavior. Questions about sexual thoughts and fantasies cannot meet the falsifiability requirements of science, and, will reduce the use of the polygraph from a scientific test to a bogus-pipeline or interrogation prop. Although potentially gratifying for a short time, polygraph questions about sexual thoughts and fantasies not expressed in behavior may ultimately lead to the devaluing and replacement of the polygraph with other emerging methods for scientific lie detection and credibility assessment. An argument is made for the use of behaviorally descriptive test stimuli that will help to ensure that polygraph testing conforms to the falsifiability requirements of science and comports with the established knowledge-base on the scientific polygraph test.

Questionable Questions

R1: Have you deliberately concealed any sexual thoughts or fantasies from your therapist?[2]

R2: Have you deliberately hidden any sexual thoughts or fantasies from your therapist?

Central to the effectiveness of any test are the test stimuli. Test stimuli during polygraph testing take the form of questions, intended to prompt the examinee to choose between the options of deception and truth-telling with a verbal answer: *yes* or *no*. According to the Model Policy for Post Conviction Sex Offender Testing (PCSOT; American Polygraph Association, 2009) and other publications,

1 Raymond Nelson is a psychotherapist, behavior scientist, trainer, and polygraph examiner who has conducted several thousand polygraph examinations. He has expertise in working with perpetrators and victims of sexual crimes and other abuse and violence. Mr. Nelson has expertise in statistics and data analysis and is one of the developers of the OSS-3 scoring algorithm and the Empirical Scoring System. He is a researcher for Lafayette Instrument Company (LIC), a developer and manufacturer of polygraph and life-science technologies. Mr. Nelson is a past-President of the American Polygraph Association (APA), currently serving as an elected Director. Mr. Nelson teaches and lectures frequently throughout the United States and internationally and has published numerous studies and papers on all aspects of the polygraph testing, including the psychological and physiological basis, test data analysis, faking/countermeasures, interviewing and question formation and test target selection. Mr. Nelson has been involved in policy development at the local, state, national and international levels in both polygraph and psychology, and has testified as an expert witness in court cases in municipal, district, appellate, superior and supreme courts. Mr. Nelson is also the academic director of the International Polygraph Training Center (IPTC). There are no proprietary or commercial interests and no conflicts of interest associated with the content of this publication. The views and opinions expressed in this publication are those of the author and not necessarily those of the APA, LIC or IPTC. Mr. Nelson can be reached at raymond.nelson@gmail.com.

2 These questions were taken verbatim, with the removal of the therapist name, from a case that was discussed in a recent courtroom proceeding.

the basic requirements for effective polygraph relevant questions are that the questions are: closed, descriptive of the examinee's behavioral involvement in an issue of concern, simple, direct, and easily understood, time-delimited, free of assumptions of guilt, free of idiosyncratic jargon and legal terms that are not easily understood by the examinee and others. Another requirement is that questions are free of references to mental state or motivational terminology except to the extent that memory or sexual motivation may be the target of the investigation following an admission of the behavior. It is also important that the questions are structurally and linguistically balanced so that they require similar attention and effort to understand. And finally, questions should be formulated in a manner that does not allow an examinee to rationalize one's involvement in a behavior or truthfulness about involvement in a behavior.

The relevant questions above are questionable for several reasons. First, they portend to investigate *sexual thoughts and fantasies* that have not manifested in actual behavior such as masturbation or sexual contact with another person. Second, they are unbounded in time (i.e., the time-delimitation refers to one's entire lifetime of *sexual thoughts and fantasies*). Although the time of reference may have been discussed during the pre-test interview, best practice is that the question clearly describes both the behavior and time period of concern. Reliance on pretest explanation is a form of logic that assumes that *the-examinee-knew-what-I-meant-even-though-the-question-did-not-articulate-it.* Instead, it should be assumed that all polygraph examinees will be confused if the questions are not unambiguous.

Some may attempt to argue that the verbs "*concealing*" or "*hiding*" (or any other synonyms for *lies* and *deception*) are the behavior that is under investigation with these questionable questions. This is incorrect. The problem with this argument is that it would require that the polygraph can measure or detect lies *per se*. It is preferable if polygraph questions make use of action verbs that describe observable behavior instead of passive or unobservable activity.

Polygraph Does Not Measure Lies *Per Se*

It is well known that the polygraph test does not measure lies *per se* (Nelson, 2014). All scientific tests, including the polygraph test, are intended to quantify things that cannot be subject to deterministic observation or direct physical measurement. Tests work as a function of proxy or substitute data sources for which observable and recordable response have been shown to have some statistical relationship or correlation, with the phenomena of interest, though the test data and test result are not themselves the phenomena of interest. Test data enable us to make probabilistic inferences to quantify the phenomena of interest. Probabilistic inferences can be used to support categorical conclusions.

The polygraph test, like all tests, is ultimately a matter of stimulus and response. The analytic theory of the polygraph test is that greater changes in physiological/autonomic activity are loaded at different types of test stimuli (i.e., differential salience of the stimuli) as a function of deception or truth-telling in response to the investigation target stimuli (American Polygraph Association, 2011; Horowitz, Kircher, Honts & Raskin, 1997; Kircher & Raskin, 1988; Honts & Peterson, 1997; National Research Council, 2003; Nelson, 2016a; Offe & Offe, 2007; Office of Technology Assessment, 1983; Raskin, Kircher, Honts & Horowitz, 1988; Senter, Weatherman, Krapohl & Horvath, 2010). Tests are conducted as a simple matter of presenting a test stimulus and then observing, recording and quantifying the response.

To increase the precision and effectiveness of test results, tests often use multiple types of data and multiple iterations or presentations of the test stimuli. This can be observed in the polygraph test using multiple types of recording sensors and multiple repetitions of the sequence of test questions. Test data from several sensors and several iterations of the test stimuli are aggregated together using structured procedures – algorithms – for which the combined data can be mathematically optimized to achieve a greater diagnostic correlation or coefficient than the data from any individual sensor. Tests data are interpreted by comparing the aggregated data to a reference model that helps us to choose the best from different possible conclusions while optimizing our practical goals in the form of test sensitivity and specificity, or false-positive and false-negative error rates.

Ultimately, all test results – including categorical conclusions – are probability statements[3]. Scientific tests are expected to quantify those probabilities as a basis of information upon which to base a conclusion. The questionable questions above encourage referring professionals and others to make the naive and dangerous assumption that the polygraph can somehow measure lies. It cannot.

It Is Not Humanly Possible to Report All *Sexual Thoughts and Fantasies*

Sexual thoughts and fantasies can occur rather frequently, with some estimates suggesting they may occur several times per day for both males and females[4]. Notwithstanding the important concerns about whether it is clinically or therapeutically desirable or necessary, in any evidence-based treatment modality, to report all *sexual thoughts and fantasies* to one's therapist, we can start by considering that an individual therapy session is exactly 50 minutes by tradition. A week includes 168 hours or 10,080 minutes.

If a person attempted report all *sexual thoughts and fantasies* it would usurp the therapy time and would prevent discussion of other important therapy issues. Neither the use of paper-and-pencil reporting procedures nor technology-based reporting protocols will rectify the fact that one cannot possibly report *all* sexual thoughts or fantasies to one's therapist. There is simply not enough time. If it were therapeutically necessary to report all sexual thoughts or fantasies to one's therapist, then any progress in sex offense specific therapy would be impossible. Therapy would be pointless. The fact that not all reports can be reviewed in detail serves to illustrate that it is not therapeutically necessary or desirable for one to report all *sexual thoughts and fantasies* in therapy. In reality, some thoughtful choices will always be made about which *sexual thoughts and fantasies* to discuss, and what level of detail to discuss them. Is this not deliberate? It is.

Wise clinicians will understand that over-intrusion into non-deviant *sexual thoughts and fantasies* (i.e., those that are within normal limits) may be counter-therapeutic. Equally important, the presence of a third-party observer in the form of a polygraph examiner, may impair the therapeutic discussion of sexual thoughts and fantasies. This type of a delicate and careful discussion must occur in a safe context so that a therapist can work to rectify distorted thoughts and feels about one's *sexual thoughts and fantasies.* Polygraph examiners are concerned primarily with obtaining information and may resort to the use of psychological manipulation of an examinee's thoughts, feelings and perceptions to accomplish this. This can result in iatrogenic effects. It would be preferable to investigate only the underreporting of masturbatory fantasies indicative of sexual deviancy. Moreover, it may be preferable to limit the use of polygraph testing of sexual deviant masturbation fantasies to examinees who are not presently engaged in therapeutic work on *sexual thoughts and fantasies.*

Some may argue that the verbs "*concealed*" and "*hidden*" are the behavioral action targets of these questions. This is incorrect. Words such as "concealed" and "hidden" are

3 An anonymous reviewer pointed out that the word "test" is sometimes used more broadly. The author generally agrees. For example, blood tests such as those for hemoglobin, sodium or potassium are quantifying these things directly, though these may be measurements. To the extent that blood levels are dynamic and situational, any measurement of these levels is still an estimate of the actual level based on data from blood sample. Another example: x-rays and ECG are not actually tests but are procedures for obtaining recorded imaging formation that then requires clinical interpretation from a knowledgeable expert.

4 A recent self-report survey by Fisher, Moore & Pittinger (2012) placed the median at 19 thoughts per day for men and 10 for women. There was wide ranging variability for both groups, 1 to 388 for male participants and 1 to 140 for female participants. Other data by Alexander & Fisher (2010) using bogus-pipeline methods found only negligible differences for males and females, suggesting that self-reported differences are influenced by social factors.

synonymous with "*lying*" and "*deception*" when used by field polygraphists. Because they do not refer to behavior with any other action verbs, these questionable-questions about "deliberately concealing or hiding any *sexual thoughts and fantasies* from [one's] therapist" place the effectiveness of the polygraph squarely on its ability to measure or detect lies *per se*[5]. If not contingent on the ability to directly measure or detect lies, these questionable-questions place the use of the polygraph test in the unscientific realm of mind reading.

The Polygraph Is Not a Mind Reading Tool

It will be important for all professionals to remain humbly aware that – for the present – it is humanly impossible for us to ever know everything about another person's *sexual thoughts or fantasies*. All scientific test results are probability statements, including when probability results are simplified to categorical results. The purpose of any scientific test is to quantify some interesting phenomena that cannot be subject to deterministic observation or physical measurement. Neither the polygraph nor any scientific test can read minds.

Returning to the questionable question examples above, involving *sexual thoughts and fantasies*, we are reminded that words such as "*concealed*" or "*hidden*" are synonyms for "*lied*" or "*falsified*" and "*withheld*" and that the polygraph machine cannot measure or detect lies *per se*. We should also be reminded that the psychological basis of responses to polygraph stimuli is thought to involve several processes including attention, cognition, emotion, and behavioral conditioning. Recorded test data are a combination of physiological proxies that have been shown to vary significantly in response to different types of test stimuli as a function of deception and truth-telling in response to relevant target stimuli.

Interpretation of the practical meaning of the test data depends on both the sensitivity and specificity of the proxy signals to deception and truth-telling, the alpha boundaries at which differential salience of the test stimuli – expressed as the loading of greater changes in physiological activity – will be regarded as statistically significant, and the prior probability of deception and truth-telling. Overarching any probabilistic and categorical inference of deception and truth-telling, is the fact that all probabilistic inferences about deception and truth-telling are based on mathematical and statistical combination of data points that are correlated with the difference between deception and truth-telling when data are elicited and recorded using standardized procedures that conform to the requirements of science.

Recorded test data are themselves neither truth nor deception. Similarly, recorded test data are not themselves the behavioral issue under investigation. And finally, the recorded data are not themselves, and cannot be taken to be, the examinee's sexual thoughts and fantasies. All test data are a form proxy that are correlated with, and so they can be used to make probabilistic calculations of, the issue of concern. Our ability to make probabilistic calculations is contingent upon our ability to study the statistical relationship between the data and the phenomena of interest – in this case, sexual thoughts and fantasies.

Without the ability to read minds, we will face a difficult or impossible challenge in finding some suitable external criterion for with which to calculate the statistical relationships between the recorded data and one's actual *sexual thoughts and fantasies (i.e.,* some criterion that does not depend on self-report information that will be subject to the same underreporting problem as these polygraph questions). We simply cannot know *everything* about a person's *sexual thoughts and fantasies* – including when a person is reported as "passing" polygraph questions about *deliberately concealing sexual thoughts and fantasies.*

5 Use of the words "concealed" and "hidden" or other similar words such as "withheld" along with other behavioral action verbs may or may not relieve the assumption here. For example: "Did you masturbate to sexual thoughts of violence any times that you have concealed from your therapist?"

Science Requires Falsifiability

Falsifiability means that there is some conceivable means to determine if the data and conclusion are incorrect. Neither *thoughts* nor *fantasies* – when these are not expressed in behavior – can be subject to deterministic observation or to direct physical measurement. Without the ability read minds, there is no conceivable way to obtain any extra-polygraphic or independent evidence such that we could ever know ground-truth. Use of *sexual thoughts* and *fantasies* as a polygraph investigation target is unfalsifiable and therefore unscientific. Use of these questionable-questions reduces the polygraph from a scientific test to a bogus-pipeline prop[6].

Some types of study are difficult or prohibitive due to cost or under-developed study methodologies. Where needed scientific study it is seen as temporarily prohibitive, science allows us to remain open to the potential scientific value of an idea – until some future time when scientific study is more likely to occur. When an idea has not been studied it is simply regarded as an *unstudied hypothesis*. Where evidence has emerged to controvert an idea – where the evidence shows an idea to be false – then the idea is referred to as a *false hypothesis*. When there is no conceivable way to obtain any external evidence that is completely independent of a test result – when the idea is conceptually *unfalsifiable* – then the idea exists outside the realm of science.

To exist in the realm of science polygraph target questions must offer some conceivable way of obtaining or locating – even if not immediately possible – some independent information to confirm or refute our conclusions as correct or incorrect. Only in this way can we begin to claim any scientific knowledge about the potential effectiveness of the polygraph test at discriminating truth and deception. A simple way to ensure that polygraph questions remain in the realm of science is to require that polygraph questions describe a behavior that can conceivably be verified/falsified as a function of some independent observable evidence.

With respect to polygraph questions about *sexual thoughts and fantasies*, some will attempt to resolve un-falsifiability problem with pragmatism, stating that confessions are enough to support the use of these questions. Although confessions are sufficient to resolve questions about individual cases, they are insufficient to answer questions about the scientific validity of the polygraph and the scientific principles of the polygraph. Confessions are insufficient as a scientific criterion – to validate the polygraph – because confessions are not independent of the polygraph if they are obtained in response to actions that are prompted by the polygraph result. We have observed, in some polygraph research, what appears in hindsight to have been the systematic exclusion of unconfirmed error cases (both false-positive and false-negative errors) because erroneous polygraph results will be much less likely to result in confirmation[7]. Datasets of confession confirmed cases have produced accuracy estimates that are so close to perfection that they cannot be viewed as representative of the kind of test accuracy and effectiveness to expect from real-world situations that may undoubtedly include unconfirmed errors.

Polygraph questions about *sexual*

6 Jones & Sigall (1971) showed that false information could be reduced in self-report surveys – initially involving racial prejudice – through the use of a fake lie-detector (i.e., a bogus pipeline) because people tend to reduce false answers to avoid incongruity when they believe their answer will be judged by a machine. Whereas bogus-pipeline research is subject to some ethical considerations or controversy because of its reliance on deception with the research participants, it is not regarded as unscientific because no actual attributions or expectations of test effectiveness are made about the bogus-pipeline results.

7 False-positive errors will not be discovered without the correct identification of a true-positive and this is not always possible. The result, when not all false-positive errors can be identified and included in a dataset, is a study sample that is non-representative and underestimates the proportion of true-positive errors. Similarly, false-negative error cases may be systematically excluded from research datasets, without additional evidence, unless an examinee notifies an examiner after producing a false-negative error. It is easy to imagine that some examinees will not be motivated additional to provide this information after such an error has occurred.

thoughts and *fantasies* – independent of behavior – cannot be falsified through information that is completely independent of the polygraph. Verification of thoughts and fantasies that do not manifest in actual behavior will depend on additional information from the examinee, and the examinee's willingness to reveal information is a function of the polygraph (it is a practical goal of polygraph testing to motivate examinees to disclose information before and after a polygraph)[8]. It is therefore unscientific to attempt to use the polygraph as a test of one's *sexual thoughts* and *fantasies* when these are not expressed in behavior. Because they are un-falsifiable, polygraph questions about *sexual thoughts and fantasies* are indicative of the use of the polygraph as a bogus-pipeline or as an interrogation prop. Any reliance on the test results from polygraph questions about *deliberately concealing sexual thoughts and fantasies* from one's therapist is outside the realm of science.

Concealment and Hiding in the Polygraph Context Are Inherently Deliberate Actions

During the polygraph pretest interview, polygraph examinees are fully informed of the target issues under investigation, and they are advised to report and discuss any information related to the target issues. Any information that remains concealed or hidden through some non-deliberate action could only exist through some mechanism involving the complete and total psychological repression of information beyond any conscious recollection[9]. Although highly unlikely, complete and total repression would be expected to mitigate any deceptive responses to relevant target stimuli. Use of the term "*deliberate*" is therefore redundant and un-necessary. If not redundant and un-necessary, this use of the term "*deliberate*" would seem to endorse the notion that polygraph can somehow test for information that has been subject to complete and total repression – such that it would be non-deliberately concealed or hidden. At this time the polygraph has not been validated as capable of testing and uncovering repressed memories, nor would there be any reason to expect to provide this capability.

All the issues surrounding the use of the term "*deliberate*" will also apply to synonymous terms such as "*intentionally*," "*willfully*," "*knowingly*," or "*consciously*," and other terms. The only way to proceed with the notion that polygraph testing can somehow quantify or discriminate "*deliberate concealment*" from either "*truth*" or "*non-deliberate concealment*" is if the polygraph can read minds. It cannot[10].

Polygraph Is Not a Test of *Mens Rea*

There are additional problems with these questionable questions when considering whether we want to endorse the use of the polygraph as a test of *mens rea* (i.e., the degree of culpability). *Mens rea* – the degree of culpability for one's behavior – is a legal consideration that is ultimately decided by the courts (Martin, 2003). Polygraph is concerned only with determining the truth about behavior – what the examinee has or has not done. Use of the polygraph as a test for *mens rea* would require that we endorse the notion that a person could *conceal* or *hide* information for non-deliberate reasons, for which they are neither culpable nor responsible for the choice to do so. The polygraph test is neither intended to be nor capable of being a test of *mens rea* or culpability. Instead, polygraph examiners have

8 To the extent that past masturbation behaviors present substantial practical barriers to falsification and may be subject to this same concern. However, falsifiability does not imply that we can immediately or easily falsify or verify every case incident. It implies only that some conceivable means exists to develop independent information. The author therefore does not view masturbation questions as unfalsifiable.

9 Repression of psychological information beyond consciousness is a topic that has been subject to some reworking of professional viewpoints in response to emerging research information (Rofé, 2008).

10 Inclusion of the term "deliberate" or its synonyms may have one other potential use. It can serve as a posttest interrogation wedge for examinees who show statistically significant reactions to relevant test stimuli, and who may have additional information to disclose. In this case, the word "deliberate" or its synonyms, though not necessary to support the logic of the test question, may serve to soften the entry into the posttest interrogation by pretending socially that the concealment or hiding of information may have been un-intentional or non-deliberate.

historically taken a pragmatic approach which holds that careful review and discussion of the target issues under investigation are enough to ensure conscious awareness, and deliberate choice, as to deception and truth-telling in response to relevant target stimuli.

Suggested Questions

To improve the effectiveness of PCSOT examinations, The American Polygraph Association (2009) Model Policy for PCSOT includes suggested targets for different types of PCSOT examinations, including Maintenance Exams, section 8.4.2.3, which address compliance with the requirements of sex offense specific treatment and supervision programs. Section 8.4.2.3 does not include *sexual thoughts* or *fantasies* – which would include both deviant and non-deviant thoughts and fantasies – as a recommended target issue for PCSOT Maintenance Exams. Recommended target issues in section 8.4.2.3 do include behavioral *masturbation to deviant fantasies*.

If professionals wish to use the polygraph to investigate issues of sexual deviancy, and if the polygraph is more than simply an interrogation prop, then the following questions are more consistent with the use of the polygraph as scientific test of deception or credibility.

R1: During the past three months have you masturbated to any thoughts or fantasies of violence?

R2: During the past three months have you masturbated to any thoughts or fantasies of prepubescent children?

Because of the large potential for differences in personal value judgments around the appropriateness or inappropriateness of different types of sexual behavior, the notion of deviancy in this context will be most useful if it is restricted to sexual acts involving underage persons who cannot legally consent to sexual activity, and sexual acts involving violent/non-consensual sexual acts with persons who do not wish to consent to sexual activities. These suggested questions better illustrate the ongoing presence or absence of observable behavioral problems involving sexual deviancy. These suggested questions are time-delimited to a shorter period, making their interpretation and use more straightforward towards attempting to determine the presence or persistence of sexual deviancy among convicted sex offenders whose offenses already indicate problems with sexual deviancy. These suggest questions are sufficiently behaviorally descriptive that there are conceivable ways to obtain extra-polygraph evidence – completely un-influenced by the polygraph results – to study the test effectiveness at determining involvement or non-involvement in the behavior.

Both *masturbation* and *sexual fantasies* are somewhat awkward and difficult discussion topics in therapy and in the polygraph test. It is tempting for people to communicate in slang terms that may or may not be correctly understood, and for which there is a corresponding tendency to engage in incompletely detailed discussion of sexual behaviors. There is also some potential for shame and judgment when discussing sexual fantasy and masturbation. The result of these complication is that there is a large potential for confusion when discussing these topics. To improve the clarity and usefulness of information from discussions about masturbation, sexual fantasies, and range of other sexual matters, Section 6 of the American Polygraph Association (2009) Model Policy for PCSOT provides *operational definitions* (i.e., behaviorally descriptive definitions) that attempt to answer the practical question: *what does it look like when someone does that?* Items N. and O. of Section 6 regarding sexual fantasy and masturbation are following:

N. *Sexual fantasy/erotic fantasy*: refers to a deliberate thought or patterns of thoughts, often in the form of mental imagery, with the goal of creating or enhancing sexual arousal or sexual feelings. Sexual fantasy can be a developed or spontaneous story, or a quick mental flash of sexual imagery, and may be voluntary or intrusive/involuntary.[11]

O. *Masturbation*: refers to sexual stimulation of one's genitals, often, though not always, to the point of orgasm. Stimulation can be over or under clothing, either manually or through other types of bodily contact, through the use of objects or devices, or through a combination of these methods. Although masturbation with a partner is not uncommon, masturbation for the purpose of this Model Policy refers to self-masturbation.

Sexual Deviancy in PCSOT

Sexual deviancy has been associated with increased risk for sexual recidivism (Hanson & Bussiere, 1998; Hanson & Harris, 1998; Hanson & Morton-Bourgon, 2005). The simplest form of information that can be indicative of sexual deviance may be self-report, such as when one discloses the details of one's deviant sexual behavior to a therapist. A limitation to the value of self-report information is that unlawful sexual behavior is not synonymous with sexual deviancy. Another limitation is that social stigma, personal shame, and potential legal consequences may inhibit self-reporting. Because self-reported information is often incomplete, an absence of self-reported sexual deviancy does not automatically indicate an absence of sexual deviancy. It is for this reason that professionals have sought to develop ways to test for sexual deviancy.

Phallometric testing (Freund, 1991)[12], in which sensors record changes in penile tumescence or circumference in response to various types of sexually themed stimuli, has been used to evaluate sexual deviancy. These tests use sexual arousal to deviant stimuli as an operational, observable, and recordable proxy for the broader construct of sexual deviancy. Phallometric results have been correlated with increased sexual recidivism risk (Hanson & Bussiere, 1998; Hanson & Morton-Bourgon, 2005). However, significant findings only involve stimuli involving children. Phallometric indices for stimuli involving sexual violence are not significant. Also, Hanson & Harris (1998) found that phallometric differences were not significant for sexual recidivists and non-recidivists as a dynamic indicator of risk, suggesting that pre-treatment phallometric assessment may be more diagnostic than post-treatment assessment.

Phallometric testing has known limitations including imperfect test sensitivity and specificity and potential vulnerabilities to some forms of faking (Freund, 1963; Freund, Watson & Rienzo, 1988, 1991; Laws & Holmen, 1988; Rubin & Henson, 1975; Wilson, 1998). The limitations of sexual deviancy information towards recidivism risk estimation are illustrated by the fact that while phallometric indices for pedophilia have been found to be correlated with recidivism, phallometric indices for rape and violence indices have not been significant. It is also important to recognize that while blood flow to the penis may be a useable proxy for sexual arousal, it is not synonymous with sexual arousal. Penile arousal can occur for a variety of reasons including sexual ideation, tactile stimulation of genitalia, parasympathetic withdrawal during sleep or deep relaxation and other causes. Young males have been known to experience seemingly spontaneous penile erections in response to wide ranging stimuli. For this reason, phallometric testing of juveniles is generally not recommended (Clift, Rajlic & Gretton, 2009). An additional limitation to phallometric testing is that it is somewhat invasive.

11 An anonymous reviewer pointed out that this definition is internally inconsistent because it first defines sexual fantasies as deliberate and then explained that they may occur involuntarily, and defines sexual fantasies as directed to the goal of arousal while unbidden sexual fantasies triggered by an unexpected external stimulus may have no intended goal.

12 Phallometric testing of sexual arousal was developed in post-war Czechoslovakia (now peacefully separated into the Czech Republic and Slovakia since 1993) after the failure of psychoanalytic methods and unstructured professional judgment at discriminating homosexuality from heterosexuality among young adults who may have been claiming to be homosexual to avoid conscription. Kurt Freund later escaped to Canada and continued working on phallometric testing as an assessment instrument for convicted sex offenders.

Sexual interest, in which visual reactions are recorded measured in response to stimuli representing various sexual themes, has also been suggested as a measurable proxy for sexual deviancy (Abel, Huffman, Warberg & Holland, 1998; Abel, Jordan, Hand, Holland & Phipps, 2001; Gray & Plaud, 2005; Harris, Rice, Quinsey & Chaplin, 1996; Letourneau, 2002) among child molesters. Efforts to discriminate violent sex offenders have been less successful. Also, not all convicted sex offenders will have deviant sexual arousal, interest or preferences, and some non-offending persons may have some deviant interest or preference that is no expressed in abusive or unlawful behaviors. Deviant sexual arousal, interest or preferences is itself insufficient to predict sex offender recidivism but can be used together with other data points to construct structured and actuarial risk prediction measures that have been found to discriminate recidivists from non-recidivists. Finally, and in general, the identification of sexual deviancy may contribute to increased risk estimates for sexual recidivism, but the absence of information on sexual deviancy does not equate directly with low risk.

Polygraph questions about sexual deviancy have also been suggested as a proxy for sexual deviancy (Odum, Busby & Nelson, 2016), though these authors also note that the use of the polygraph to test fantasies not connected with behavior is outside the scope of existing polygraph studies. Nelson (2016b) also cautioned that testing of fantasies not connected with behavior is outside the scope of presently established practice recommendation of the American Polygraph Association. There is presently no published literature describing the use or effectiveness of polygraphic information as an indicator of sexual deviancy or sexual recidivism risk. Polygraph is a test of credibility, referred to as a lie-detector as a matter of convenience, for which test results are a probabilistic measurement intended to support a categorical conclusion (Nelson, 2014; 2015). It is unknown whether the developers of structured and actuarial sex offense recidivism measures (Boer, Hart, Kropp & Webster, 1997; Duwe & Freske, 2012; Epperson *et. al.*, 2005; Epperson Kaul & Goldman, 2003; Harris *et al.*, 2003; Hanson, 1997; Hanson & Harris, 2000; Hanson, Harris, Scott & Helmus, 2007; Hanson & Thornton, 1999; 2000; 2003; Quinsey, Harris, Rice, & Cormier, 2006; Rice & Harris, 1997) would support the unstudied drop-in use of polygraphic information as a proxy for other information indicative of sexual deviancy and sexual recidivism. What is known is that none of the presently available structured or actuarial risk measures includes polygraphic information as a content item.

Information from the polygraph is used for its clinical value. Although highly useful in the investigation context, clinical and actuarial clinical use of information obtained through coercion may, for ethical and therapeutic reasons, be more circumspect. Meta analytic research has begun to reveal that while voluntary participation in treatment produces significant desirable effects regardless of the setting, coerced treatment has been found to be ineffective, particularly in custodial settings (Parhar, Wormith, Derkzen & Beauregard, 2008). Discussion and debate exist in the clinical literature regarding the use of coercion and the importance of maintaining and adhering to clinical ethics (Glaser, 2010; Prescott & Levinson, 2010; Ward, 2010). Although self-reported information is known to be incomplete, information obtained through non-manipulative interviewing and testing methods can be assumed to be less problematic in terms of its authenticity, and clinical and actuarial value.

Clinicians who wish to remain within the bounds of evidence-based practice will want to ensure that the information used for assessment and treatment is obtained through activities that comport with clinical and therapeutic values for humane and ethical treatment of mental health patients. Professionals who use the polygraph to develop self-report or testing information on sexual deviancy can ensure the integrity and usability of the resulting information by using interviewing and testing methods that do not rely upon psychological manipulation and distorted cognitions to obtain information. Although perhaps useful in other contexts, information that is obtained through high-pressure interactions or psychological manipulation may be regarded as insufficient as a basis for clinical work or diagnostic conclusions about sexual deviancy and may serve to interfere with the therapeutic alliance for which emerging evidence suggests plays a role in recidivism risk (Blasko & Jeglic, 2016). Regardless of how information about sexual deviancy is obtained, it is important to remember that DSM-V (American Psychiatric Association, 2013) differentiates *paraphilias*, which require no psychological or psychiatric treatment, and *paraphilic disorders* which cause social or psychological or distress and therefore require treatment.

Summary

Polygraph questions about *deliberately concealing any sexual thoughts and fantasies* from one's therapist are questionable for several reasons. These questions are without a specified time of reference and appear to refer to a person's entire lifetime. Without reference to the behavioral expression of one's sexual thoughts or fantasies, these questions are unfalsifiable and are therefore outside the realm of science, scientific testing, and the scientific use of the polygraph test. These questions rest on a series of troublesome notions about the polygraph test, beginning with the notion that the polygraph can discriminate deliberate from non-deliberate concealment of information. Other troublesome implications are that the polygraph can function as a mind-reading device, or that the polygraph can serve as a test of *mens rea*, or that the polygraph can measure or detect lies *per se*. All of these are inconsistent with reality and inconsistent with scientific polygraph and credibility assessment testing. Use of these questionable questions depends on rationalizations that compromise the integrity of scientific polygraph testing. Equally important, information from polygraph tests of one's *sexual thoughts or fantasies* is of questionable therapeutic and diagnostic value. Finally, these questionable questions also fail to meet the basic requirements for relevant questions, as described in published field practice policies within the polygraph profession.

Use of the word "*deliberate*" introduces the dubious implication that one can conceal information non-deliberately, and in doing so produce non-deceptive polygraph results through some mechanism of complete and total psychological repression of the sought-after information (i.e., outside of any conscious awareness). For both practical and linguistic purposes, use of the word "*deliberate*" is redundant. Though its inclusion may be associated with the use of the polygraph as an interrogation tool, to stage the introduction of a post test interrogation, it adds nothing to the logic or information value of the stimuli. Equally concerning, there may be some who might enjoy it if we were to endorse the notion that one may be regarded as truthful if others can be convinced that one has conceal information only non-deliberately through some process involving psychological repression.

Can one completely repress information, beyond any conscious awareness, about sexual thoughts or fantasies? Can the polygraph test for deception and truth-telling around repressed information? Is it possible to be deceptive or truthful about one's *sexual thoughts and fantasies* if they are not expressed in behavior? Can statements about *sexual thoughts or fantasies*, not expressed in behavior, be epistemologically true or false? More simply, what kinds of things can be true? And is it possible to account for one's reporting or concealment of *sexual thoughts and feelings* using a polygraph test? Accounting for one's thoughts is a tricky proposition both philosophically and scientifically. As an example, consider the following statement: "*I did not have a thought about a yellow Volkswagen.*" Or: "*Don't think about a yellow Volkswagen.*" Merely reading, thinking, hearing or stating the sentence involves a thought about a yellow Volkswagen (i.e., the recursive thought about not having thoughts about a yellow Volkswagen)[13]. With this awareness, it will be important to carefully consider whether we attempt to endorse the notion that the completeness of one's reporting of *sexual thoughts and fantasies* can be accounted for as deceptive or

13 This is a variant of the *white bear problem* from *ironic process theory*. The names stem from the fact that it was first described by Fyodor Dostoyevsky in a publication from 1863 (Winter Notes on Summer Impressions). Ironic process theory holds that deliberate attempts to suppress certain thoughts will make them more likely to surface. Another variation of this theory is a game for children, the *white bear story*, whereby we tell children to think of a white bear and hold up their hands, and then put their hands down only when they *stop* thinking about a white bear.

truthful via polygraph testing. As a practical matter, statements and questions about physical phenomena (i.e., physical things and physical events) are more readily falsifiable and more easily conform to intuitively useful notions about truth and deception. Polygraph questions about behavior will also provide clearer intuitive and practical meaning for sex offense specific treatment providers and risk evaluators.

Polygraph questions about *sexual thoughts and fantasies* – not connected with any behavior – will fail to serve the needs of sex offender treatment and supervision program if they are not connected with behavior, are un-constrained as to sexual deviancy, unbounded as to the period of the examinee's lifetime, and unrelated to sex offense recidivism risk. These questionable questions are so broad and un-bounded that it would be impossible for a person to ever completely report a lifetime of *sexual thoughts and fantasies* to a therapist. Some thoughtful choices must me made about what to report and what not to report. The notion that a person could fully report all of one's *sexual thoughts and fantasies,* or that any professional could ever fully know all of an examinee's *sexual thoughts and fantasies*, is not merely unscientific – it crosses the boundary into magical thinking. These questions serve only to reduce both the polygraph and sex offender treatment processes to attempted mind reading. They denigrate the treatment and supervision process to a counter-therapeutic shell-game that may ultimately undermine the therapeutic alliance and trust towards professionals whose objectives are to help convicted sex offenders learning to live safer and health lives.

One of the most important aspects of effective relevant questions is the requirement that they are descriptive of the examinee's possible involvement in a behaviorally for which the examinee will know the truth about his or her involvement. This requirement for behaviorally descriptive questions ensures that the relevant questions are *falsifiable.* Falsifiability means that there is some conceivable extra-polygraphic or independent means (i.e., for which it is inconceivable that that the other information could in any way be influenced by the test result) to determine if the data and conclusion are correct or incorrect. Falsifiability does not require that we immediately have access to the independent information; it means only that some conceivable way exists to obtain such information. Questions about *sexual thoughts and fantasies* – not connected with any behavior – are unfalsifiable because we cannot read a person's mind. A requirement for falsifiability ensures that the polygraph is used within the boundaries of science, and a requirement for behaviorally description questions will help to ensure that polygraph questions are likely to provide useful interpretable meaning regardless of whether the answer is deceptive or truthful. Questions about *deliberately concealing any sexual thoughts and fantasies*, independent of behavior, are associated with the use of the polygraph as a bogus-pipeline or interrogation tool instead of as a scientific test.

Questions about *deliberately concealing sexual thoughts and fantasies* cannot provide useable intuitive meaning because it can be assumed that many people will not fully report all their *sexual thoughts and fantasies* during any unbounded time period. In other words, if we attempt to interpret test results as categorically negative for the presence of deliberately concealed *sexual thoughts and fantasies*, then it cannot be reasonable assumed when a person has reported *all sexual thoughts or fantasies.* Truthful polygraph test results, concerning questions about deliberate concealing of *sexual thoughts and fantasies*, are meaningless. It would be clinically unwise and naive for any therapist to ever assume that a convicted sex offender has reported *all sexual thoughts and fantasies*, or that we can somehow know everything in this area. All persons in sex offense specific treatment can be expected to make thoughtful and deliberate choices as which *sexual thoughts and fantasies* to discuss and which not to discuss. There will be additional deliberate choices about how to discuss and view those *sexual thoughts and fantasies* that are selected for discussion. Wise and experienced mental health professionals and sex offense specific treatment providers will always remain aware that we cannot humanly know everything about another persons' *sexual thoughts and fantasies.*

Fortunately, it is not clinically or therapeutically necessary, or desirable, for a person to report all *sexual thoughts and fantasies.*

All that is necessary is for a clinician and risk evaluator to obtain enough information to determine the presence, or reasonably confirm the absence, of problems involving sexual deviancy. Once identified, it is doubtful whether things like sexual deviancy, paraphilias and paraphilic disorders can ever be completely removed or eliminated. A more likely therapeutic objective is to manage and reduce the level of problems and distress that these may cause.

In contrast, polygraph questions about recent or ongoing *masturbation* to deviant *sexual thoughts and fantasies* (i.e., masturbation to *sexual thoughts and fantasies* involving violence or prepubescent children) can provide clearer intuitive meaning when a polygraph test result is indicative of deception or truth-telling. These questions provide some conceivable means of falsification. From a clinical perspective, ongoing masturbation to deviant sexual thoughts and fantasies may be an indicator of continued reinforcement of a propensity for other sexually deviant behavior or may indicate either inability or unwillingness to curtail and reduce these behaviors. A requirement for falsifiable behaviorally-descriptive questions will help to prevent the attempted use of the polygraph as a pseudoscientific mind-reading or fortune-telling device.

Presently, there is no published scientific information suggesting that sexual deviancy or dangerousness can be measured as a function of proxy data in the form of polygraph results to questions about *deliberately concealing any sexual thoughts or fantasies*. However, self-reported information about deviant sexual thoughts or fantasies may be used in risk assessment. Information that is coerced or manipulated may be of far less therapeutic and forensic value. Polygraph results – if the test results themselves are ever to be viewed as useful (i.e., not useless) – should address issues of sexual deviancy that are expressed in behavior such as masturbation, illegal forms of pornography (child pornography, revenge pornography, voyeuristic pornography, etc.) or unlawful/abusive behavior.

The polygraph test, like all tests, is ultimately a matter of stimulus and response. The test is conducted by presenting the test stimuli and then observing, recording and quantifying the response. Tests work as a function of proxy or substitute data sources for which observable and recordable response have been shown to have some statistical relationship or correlation, with the phenomena of interest, though the test data are not themselves the phenomena of interest. Studying and defining the proxy relationship requires that there is some criterion, known with reasonable certainty, with which we can calculate the relationship between the recorded data and the criterion. In the case of *sexual thoughts and fantasies* no suitable criterion exists when these thoughts and fantasies are not expressed in behavior.

Finally, it should be remembered that this paper is a position paper that is not based on empirical study. The author reminds the reader that the acceptance of any field practice without evidence may be a hazardous and surprising adventure. Surely there will much more that can be learned when time and resources are made available to address these interesting and important discussions with actual data analysis. Until then, the author invites alternative viewpoints and cautions that best practices will restrain the use to the polygraph to within the scope of questions and targets issues for which our present existing knowledge base is extensible and generalizable.

References

Abel, G. G., Huffman, J., Warberg, B. & Holland, C. L. (1998). Visual reaction time and plethysmography as measures of sexual interest in child molesters. *Sexual Abuse: A Journal of Research and Treatment, 10(2),* 81-95.

Abel G. G., Jordan A., Hand C. G., Holland L. A., Phipps A. (2001). Classification models of child molesters utilizing the Abel Assessment for sexual interest™. *Child Abuse & Neglect, 25,* 703-718.

Alexander, M. G. & Fisher, T. D. (2010). Truth and consequences: Using the bogus pipeline to examine sex differences in self-reported sexuality. *Journal of Sex Research, 40(1)*, 27-35.

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph, 40(4),* 196-305. [Electronic version] Retrieved August 20, 2012, from http://www.polygraph. org/section/research-standards-apa-publications.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.).* Arlington, VA: American Psychiatric Publishing.

American Polygraph Association (2009). *Model Policy for Post-conviction Sex Offender Testing.* [Electronic version Retrieved] 1/25/2012, from http://www.polygraph.org.

Blasko, B. L. & Jeglic, E. L. (2016). Sexual offenders' perceptions of the client-therapist relationship: The role of risk. *Sexual Abuse: A Journal of Research and Treatment, 28(4),* 271-90.

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20: Professional guidelines for assessing risk of sexual violence.* Vancouver, Canada: The Mental Health, Law, and Policy Institute, Simon Fraser University.

Clift, J., Rajlic, G., & Gretton, H. (2009). Discriminative and predictive validity of the penile plethysmograph in adolescent sex offenders. *Sexual Abuse, 21(3),* 335–362.

Duwe, G., & Freske, P. J. (2012). Using Logistic Regression Modeling to Predict Sex Offense Recidivism: The Minnesota Sex Offender Screening Tool-3 (MnSOST-3.1). *Sexual Abuse: A Journal of Research and Treatment, 24(4),* 350-377.

Epperson, D. L., Kaul, J. D. & Goldman, R. (2003). Minnesota sex offender screening tool - revised (MnSOST-R) technical paper: Development, validation, and recommended risk level cut Scores. Minnesota Department of Corrections.

Epperson, D.L., Kaul, J.D., Huot, S., Goldman, R., Hesselton, D., & Alexander, W. (2005). *Minnesota Sex Offender Screening Tool-Revised (MnSOST-R) scoring guidelines—updated.* Retrieved from: http://www.psychology.iastate.edu/~dle/ MnSOST-R Manual12-22-2005.pdf

Freund, K. (1963). A laboratory method for diagnosing predominance of homo- or hetero-erotic interest in the male. *Behaviour Research & Therapy, 1,* 85-93.

Freund, K.; Watson, R. J. (1991). Assessment of the sensitivity and specificity of a phallometric test: An update of phallometric diagnosis of pedophilia. *Psychological Assessment, 3(2),* 254–260.

Freund, K., Watson, R., & Rienzo, D. (1988). Signs of feigning in the phallometric test. *Behaviour*

Research & Therapy, 26, 105-112.

Fisher, T. D., Moore, Z. T., & Pittenger, M. J. (2012). Sex on the brain?: an examination of frequency of sexual cognitions as a function of gender, erotophilia, and social desirability. *Journal of Sex Research, 49(1),* 69-77.

Glaser, B. (2010). Sex offender programmes: New technology coping with old ethics. *Journal of Sexual Aggression, 16(3),* 261-274.

Gray, S. R., & Plaud, J. J. (2005). A comparison of the Abel Assessment for Sexual Interest and penile plethysmography in an outpatient sample of sexual offenders. *Journal of Sexual Offender Civil Commitment: Science and the Law, 1,* 1-10.

Hanson, R. K. (1997). The development of a brief actuarial risk scale for sexual offense recidivism. (User Report No. 1997-04). Ottawa: Department of the Solicitor General of Canada.

Hanson, R. K. & Bussiere, M.T. (1998). Predicting relapse: a meta-analysis of sexual offender recidivism studies. Journal of consulting and clinical psychology, 66, 348-62.

Hanson, R. K. & Harris, A. (1998). *Dynamic Predictors of Sexual Recidivism.* Department of the Solicitor General Canada.

Hanson, R. K., & Harris, G. T. (2000). *The Sex Offender Need Assessment Rating (SONAR): A Method for Measuring Change in Risk Levels.* Ottawa, ON, Canada: Department of the Solicitor General of Canada.

Hanson, R. K., Harris, A. J. R., Scott, T.-L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project (User report, Corrections research).* Ottawa, ON, Canada: Public Safety Canada. Retrieved from http://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ssssng-rsk-sxl-ffndrs/index-eng.aspx

Hanson R. K., Morton-Bourgon K. E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73,* 1154–1163.

Hanson, R. K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sex offenders. User Report 99-02.* Ottawa: Department of the Solicitor General of Canada.

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24(1),* 119-136.

Hanson, R. K., & Thornton, D. (2003). *Notes on the development of STATIC-2002. User Report 2003-01.* Ottawa, ON, Canada: Department of the Solicitor General of Canada.

Harris G. T., Rice M. E., Quinsey V. L., Chaplin T. C. (1996). Viewing time as a measure of sexual interest among child molesters and heterosexual men. *Behaviour Research and Therapy, 34,* 389-394.

Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumiere, M. L., Boer, D., & Lang, C. (2003). A multi-site comparison of actuarial risk instruments for sex offenders. *Psychological Assessment, 15,* 413-125.

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.

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.

Jones, E. & Sigall, H. (1971). The bogus pipeline: a new paradigm for measuring affect and attitude". *Psychological Bulletin, 76(5),* 349–364.

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.

Laws, D. & Holmen, M. (1978). Sexual response faking by pedophiles. *Criminal Justice & Behavior, 5,* 343-356.

Letourneau E. J. (2002). A comparison of objective measures of sexual arousal and interest: Visual reaction time and penile plethysmography. Sexual Abuse: A Journal of Research and Treatment, 10, 81-95.

Martin, E. A. (2003). *Oxford Dictionary of Law.* Oxford: Oxford University

National Research Council (2003). *The Polygraph and Lie Detection.* National Academy of Sciences.

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

Nelson, R. (2015). Polygraph testing as a single-subject scientific experiment. *APA Magazine, 48(5),* 101-106.

Nelson, R. (2016a). Scientific (analytic) theory of polygraph testing. *APA Magazine, 49(5),* 69-82.

Nelson, R. (2016b). Testing the limits of evidence-based polygraph practices. *Polygraph, 45(1),* 74-85.

Odum, J., Busby, T. & Nelson, R. (2016). Polygraph in civil commitment programs for sexually violent predators. *Polygraph, 45(1),* 57-73.

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.* [Re-printed in Polygraph, 12, 198-319.]

Parhar, K. K., Wormith, J. S., Derkzen, D. M., & Beauregard, A. M. (2008). Offender coercion in treatment: a meta-analysis of effectiveness. *Criminal Justice and Behavior, 35(9),* 1109-1135.

Prescott, D. S. & Levinson, J. S. (2010). Sex offender treatment is not punishment. *Journal of Sexual Aggression, 16(3),* 275-285.

Quinsey, V. L., Harris, G. T., Rice, M. E. & Cormier, C. A. (2006). Violent Offenders: Appraising and Managing Risk, 2nd Edition. Washington DC: American Psychological Association.

Rofé, Y. (2008). Does repression exist?: memory, pathogenic, unconscious and clinical evidence. *Review of General Psychology, 12(1),* 63–85.

Rubin, H. B. & Henson, D. E. (1975). Voluntary enhancement of penile erection. *Bulletin of the Psychonomic Society, 6,* 158-160.

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(2),* 109-117.

Raskin, D. C., Kircher, J. C., Honts, C. R. & Horowitz, S.W. (1988). Validity of control question polygraph tests in criminal investigation. *Psychophysiology, 25,* 476.

Rice, M. E. & Harris, G.T. (1997). Cross-validation and extension of the violence risk appraisal guide for child molesters and rapists. Law and human behavior, 21, 231-41.

Ward, T. (2010). Punishment or therapy? The ethics of sexual offending treatment. *Journal of Sexual Aggression, 16(3)*, 286-295.

Wilson, R. J. (1998). Psychophysiological signs of faking in the phallometric test. *Sexual Abuse: A Journal of Research and Treatment, 10(2),* 113-126.

Multinomial Reference Distributions for Three-Position Scores of Comparison Question Polygraph Examinations

Raymond Nelson

Abstract

Multinomial theoretical reference distributions, calculated under the analytic theory of the comparison question test (CQT), are shown for three-position numerical scores for three to five repetitions of test question sequences that include two to four relevant questions. Multinomial reference distributions were previously published for the Empirical Scoring System (ESS) but have not previously been made available for three-position scores. Three-position scores are a foundation from which ESS scores are derived. However three-position scores involve stronger reliance on naive assumptions that different recording sensors may contribute equally to test outcomes. Reference tables are shown for the traditional array of polygraph recording sensors – respiration, electrodermal, cardio – and also including the vasomotor sensor. Multinomial reference tables may be useful as a likelihood-function to obtain reproducible statistical values for three-position CQT scores, can be used to optimize numerical cut-scores for use with three-position numerical scores, and can be useful in Bayesian analysis of polygraph test data.

Introduction

Nelson (2017) calculated multinomial reference distributions for Empirical Scoring System (ESS; Nelson, Krapohl & Handler, 2008; Nelson, et. al., 2011) scores of the CQT. ESS scores are derived from three-position scores (Bradley & Janisse, 1981; Department of Defense, 2006b; van Herk, 1991). Three-position scores differ from ESS scores in that three-position scores are premised on a strong (i.e., deliberate) naive assumption that little is known about the relative importance or contribution of the different recording sensors. Naive assumptions are commonly used in Bayesian analysis (Berger, 1985; Box & Tiao, 1973; Jaynes, 1986; 2003; Lee, 2004; Rubin, Gelman, Carlin, & Stern, 2003) and machine learning (Hand & Yu, 2001; Russell, & Norvig, 2003) because they can permit us to develop simple and effective solutions to complex and difficult problems. Although empirical reference distributions were provided by Nelson and Handler (2015), prior to this time no published description has been available for theoretical distributions of three-position CQT scores. This project is intended to fill a gap in the knowledge base.

Multinomial reference distributions were calculated for three-position scores under the null-hypothesis to the analytic theory of the comparison question test (CQT). The analytic theory of the polygraph, as described in numerous studies on the evaluation of comparison question polygraph data (Barland, 1985; Kircher & Raskin, 1988; Krapohl & McManus, 1999; Kubis, 1962; Nelson, Krapohl & Handler, 2008; Summers, 1939), holds that greater changes in physiological activity are loaded at different types of test stimuli as a function of deception or truth-telling in response to relevant or investigation target stimuli. [See Nelson (2016) for a discussion of this analytic hypothesis.] This theory does not attempt to describe the exact psychological or physiological mechanisms underlying responses during the CQT, and only expresses how the data are expected to be observed.

Part of the value of an analytic theory for the CQT is that it can be expressed in the form of a null-hypothesis. The null-hypothesis would state that the distributions polygraph scores are not loaded in any systematic way that discriminates deception or truth-telling. In this usage, our expression of the null-hypothesis is not for the purpose of testing or evaluating the null hypothesis against the alternative hypothesis but is instead an attempt to use the null-distribution to obtain a likelihood statistic that can be to classify unknown cases using Bayesian analysis. Because the individual scores can be characterized as random under the null-hypothesis to the analytic theory, the distribution of the summed three-position sensor scores under the null-hypothesis can be calculated using combinatoric math or through simulation. The distribution of CQT scores is multinomial (as opposed to binomial) because there are more than two possible values for each sensor score for iteration of the test stimuli.

Discussion

Calculation of the multinomial distribution[1] of CQT scores involves first the calculation of the multinomial distribution of sensor subtotals. The distribution for CQT total scores is then calculated as the combination of the distributions of the sensor subtotals. All calculations were completed with the R statistical computing language (R Core Team, 2018).

Multinomial Distribution of Sensor Subtotals

Calculation of the multinomial reference distribution for three-position CQT scores begins with the distribution of subtotal scores for the individual recording sensors. Polygraph question formats currently in use in field settings include the use of two, three and four relevant questions. Subtotal scores for individual relevant questions are often used when evaluating the results of multiple-issue screening tests. For this reason the multinomial sensor distributions were calculated for grand total scores with one, two, three and four relevant questions. Because published evidence (Kircher & Raskin, 1988; Senter & Dollins 2004; Senter, Dollins & Krapohl, 2004) has shown that conducting up to five, with a minimum of three, repetitions of the question sequence can increase the effectiveness of deceptive and truthful classifications while reducing the occurrence of inconclusive outcomes, multinomial reference distributions for the sensor scores were calculated for up to five iterations of the question sequence.

For each iteration of each relevant question, the distribution of three-position scores [-1, 0, +1] is expected to be random under the null hypothesis, and the scores are therefore equally likely to occur. A polygraph test with up to five repetitions of a question sequence that includes three relevant questions will have recording sensor scores that range from -15 to +15. There will be only 1 way to achieve either the maximum or minimum sensor score: every iteration of each relevant question would have to produce a +1 or -1 score. In contrast, there are numerous ways to achieve other possible sensor subtotals, and the most commonly occurring sensor subtotal score (i.e., modal value) will be 0 (zero) under the null-hypothesis. The multinomial distribution of sensor subtotals will inform us of the number of ways and statistical likelihood associated with each possible sensor score.

Reference tables for sensor subtotals are shown in Appendices A-C for CQT formats consisting of five repetitions of question sequences including two, three and four relevant questions with three-position scores. Appendix D shows the reference table for sensor subtotals with five repetitions of a single relevant question. The tables shown in Appendices A-D include the number of *ways* to achieve each possible sensor score along with the *probability mass function* (*pmf*) that describes the statistical likelihood of observing each possible score under the null hypothesis. The cumulative sum of the *ways* will be equal to n^k, where *n* number of possible scores (i.e., [-1, 0, +1]) for each iteration of each relevant question and *k* is the product of the number of relevant questions and the number of repetitions of the question sequence. The *pmf* will be used to calculate the multinomial distribution of the combined sensor scores. [Refer to Nelson (2017) for more information.]

1 A *distribution* tells us the range of possible values. Distributions can be expressed in table form, and often include the statistical likelihood associated with each value.

Multinomial Distribution of Total or Combined Sensor Scores

The multinomial distribution of total scores is the combination of the multinomial distributions for the sensor subtotals. For this reason, the distributions for sensor subtotals must be calculated before it is possible to calculate the distributions for total scores. Similar to the distributions for sensor subtotals, the multinomial distribution for total scores will include only one possible way to achieve the maximum or minimum possible score – all sensor subtotals must achieve the maximum or minimum score. The distribution of total scores will include numerous ways of achieving total scores near the modal total score of 0 (zero).

The maximum possible three-position CQT total score will be the product of the number of recording sensors and the maximum possible sensor score. Similarly, the minimum possible total score will be the product of the number of recording sensors and the minimum possible sensor score. For example: a CQT with three relevant questions using the normal array of respiration, electrodermal and cardiovascular sensors will have a distribution of total scores in the range [-45...+45]. Addition of the vasomotor recording sensor will increase the range to [-60...+60].

Appendices E-G show the multinomial reference distributions for grand total scores with CQT question sequences that include two, three and four relevant questions using the traditional array of respiration, electrodermal and cardio recording sensors. Appendix H shows the multinomial reference distribution for subtotal scores of individual RQs, including statistical corrections for multiplicity. Reference distributions are shown with the additional of the vasomotor recording sensor in Appendices I-L. Reference tables in Appendices E-L can serve as a likelihood function to obtain posterior estimates of the likelihood of deception or truth-telling using Bayesian analysis.

Reference tables in Appendices E-L include the number of ways to achieve each total score, along with the *pmf* and *cumulative distribution function* (*cdf*). Also included in these appendices is a continuity-corrected (*cdfContCor)* from which statistical estimates for CQT total scores. The continuity correction is a mathematical correction moves each value slightly further from the modal value (score = 0) and ensures that estimated statistical values always exceed the actual value. Appendices E-L also include a transformation of the *cdfContCor* to an *odds*, using the formula odds = p / (1-p)[2]. Finally the lower limit of the 1-*a*/2% interval for the odds is shown in the *oddsLL05* column using the Clopper-Pearson method (Clopper & Pearson, 1934) [See Nelson (2018a) for more information on the Clopper-Pearson interval in this usage.]

In practical use the tables in Appendices E-L can be used to determine numerical cut-points prior to polygraph testing. [Refer to Nelson (2018b) for more information on the use multinomial reference tables for the CQT.] Cut-scores for the three-position multinomial reference model are shown in Table 1 for event-specific diagnostic polygraph exam, and in Table 2 for multiple-issue screening polygraphs.

2 A proportional estimate can be re-calculated from the odds using `p = odds / (1 + odds)`.

Table 1. Multinomial cut-scores for three-position grand total scores of event-specific exams, using a one-tailed alpha = .05 for the lower limit of the Clopper-Pearson interval for positive and negative classifications (multiplicity-corrected subtotal cut-scores in parenthesis).

	2 RQs	3 RQs	4RQs
Respiration, EDA, Cardio	+2 / -2 (-5)	+2 / -2 (-6)	+2 / -2 (-8)
Respiration, EDA, Cardio, Vasomotor	+2 / -2 (-4)	+2 / -2 (-6)	+2 / -2 (-8)

Table 2. Multinomial cut-scores for three-position subtotal scores of multiple-issue exams, using a one-tailed alpha = .05 for the lower limit of the Clopper-Pearson interval. Subtotal cut-scores are without statistical correction for positive classifications and with statistical correction for negative classifications.

	2 RQs	3 RQs	4RQs
Respiration, EDA, Cardio	+2 / -3	+1 / -3	+1 / -3
Respiration, EDA, Cardio, Vasomotor	+2 / -3	+1 / -3	+1 / -3

Conclusion

The availability of multinomial theoretical reference distributions represents a substantial advance in the scientific foundation of the CQT. Theoretical reference distributions are important in both scientific research and scientific testing because they can be used to make inferences about the statistical likelihood of observed data. These multinomial tables can be used as a likelihood function for Bayesian analysis of CQT score, either manually or through automated algorithm.

Theoretical reference distributions differ from empirical reference models in that they are calculated from facts and information that can be subject to mathematical and logical proof. In contrast, the calculation of empirical reference distributions depends on the availability of a numerous empirical samples, and rigorous sampling methodology, in order to closely approximate the unknown population distribution. When both theoretical and empirical distributions are available it will be expected that observed results and conclusions should not differ substantially for the two methods.

Calculation of these distributions would have been substantially more unwieldy before computers and powerful statistical programming environments were available. It is hoped that the publication of these multinomial reference distributions may be of some use to field examiners and program managers interested in the use of evidence-based practices with the three-position CQT scores and may help facilitate additional interest in research on the objective analysis of CQT scores.

References

Barland, G. H. (1985). A method of estimating the accuracy of individual control question polygraph tests. *Proceedings of IDENTA, 1985,* 142-147.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* Springer-Verlag.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Wiley.

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

Clopper, C.; Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika. 26,* 404–413.

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

Hand, D. J. & Yu, K. (2001). Idiot's Bayes — not so stupid after all? *International Statistical Review. 69 (3),* 385–399.

Jaynes, E.T. (1986). Bayesian Methods: General Background. In Justice, J. H. *Maximum-Entropy and Bayesian Methods in Applied Statistics.* Cambridge University Press.

Jaynes E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press.

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.

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

Kubis, J. F. (1962). *Studies in Lie Detection: Computer Feasibility Considerations. RADC-TR 62-205, Contract AF 30(602)-2270.* Air Force Systems Command, U.S. Air Force, Griffiss Air Force Base. New York: Rome Air Development Center.

Lee, P. M. (2004). *Bayesian Statistics, an introduction (3rd ed.).* Wiley.

Nelson, R. (2016). Scientific (analytic) theory of polygraph testing. *APA Magazine, 49(5)*, 69-82.

Nelson, R. (2017). Multinomial reference distributions for the Empirical Scoring System. *Polygraph & Forensic Credibility Assessment, 46(2),* 81-115.

Nelson, R. (2018a). Five-minute science lesson: Clopper-Pearson credibility intervals for Bayesian analysis of multinomial polygraph scores. *APA Magazine, 51(3),* 61-70.

Nelson, R. (2018b). Guide for how to use the ESS-multinomial reference tables in four steps. *APA Magazine, 51(2),* 78-89.

Nelson, R. & Handler, M. (2015). Statistical reference distributions for comparison question polygraphs. *Polygraph, 41(1)*, 91-114.

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.

Rubin, D. B., Gelman, A., Carlin, J. B. & Stern, H. (2003). *Bayesian Data Analysis (2nd ed.).* Boca Raton: Chapman & Hall/CRC.

Senter, S. & Dollins, A. (2004). Comparison of question series and decision rules: A replication. *Polygraph, 33(4)*, 223-233.

Senter, S., Dollins, A. & Krapohl, D. (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.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Russell, S. & Norvig, P. (2003) *Artificial Intelligence: A Modern Approach (2nd ed.).* Prentice Hall.

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

Appendix A.

Multinomial Reference Distribution for Sensor Totals with 5 Repetitions of 2 Relevant Questions

Score	Ways	Pmf
-10	1	<.0001
-9	10	.0002
-8	55	.0009
-7	210	.0036
-6	615	.0104
-5	1452	.0246
-4	2850	.0483
-3	4740	.0803
-2	6765	.1146
-1	8350	.1414
0	8953	.1516
1	8350	.1414
2	6765	.1146
3	4740	.0803
4	2850	.0483
5	1452	.0246
6	615	.0104
7	210	.0036
8	55	.0009
9	10	.0002
10	1	<.0001

Appendix B.

Multinomial Reference Distribution for Sensor Totals with 5 Repetitions of 3 Relevant Questions

Score	Ways	Pmf
-15	1	<.0001
-14	15	<.0001
-13	120	<.0001
-12	665	<.0001
-11	2835	.0002
-10	9828	.0007
-9	28665	.0020
-8	71955	.0050
-7	157950	.0110
-6	306735	.0214
-5	531531	.0370
-4	827190	.0576
-3	1161615	.0810
-2	1477035	.1029
-1	1704510	.1188
0	1787607	.1246
1	1704510	.1188
2	1477035	.1029
3	1161615	.0810
4	827190	.0576
5	531531	.0370
6	306735	.0214
7	157950	.0110
8	71955	.0050
9	28665	.0020
10	9828	.0007
11	2835	.0002
12	665	<.0001
13	120	<.0001
14	15	<.0001
15	1	<.0001

Appendix C.

Multinomial Reference Distribution for Sensor Totals with 5 Repetitions of 4 Relevant Questions

Score	Ways	*Pmf*
-20	1	<.0001
-19	20	<.0001
-18	210	<.0001
-17	1520	<.0001
-16	8455	<.0001
-15	38304	<.0001
-14	146490	<.0001
-13	484500	.0001
-12	1409895	.0004
-11	3656360	.0010
-10	8533660	.0024
-9	18062160	.0052
-8	34880770	.0100
-7	61757600	.0177
-6	100640340	.0289
-5	151419816	.0434
-4	210859245	.0605
-3	272290140	.0781
-2	326527350	.0936
-1	363985680	.1044
0	377379369	.1082
1	363985680	.1044
2	326527350	.0936
3	272290140	.0781
4	210859245	.0605
5	151419816	.0434
6	100640340	.0289
7	61757600	.0177
8	34880770	.0100
9	18062160	.0052
10	8533660	.0024
11	3656360	.0010
12	1409895	.0004
13	484500	.0001
14	146490	<.0001
15	38304	<.0001
16	8455	<.0001
17	1520	<.0001
18	210	<.0001
19	20	<.0001
20	1	<.0001

Appendix D.

Multinomial Reference Distribution for Sensor Subtotals with 5 Repetitions of the Question Sequence

Score	Ways	Pmf
-5	1	.0041
-4	5	.0206
-3	15	.0617
-2	30	.1235
-1	45	.1852
0	51	.2099
1	45	.1852
2	30	.1235
3	15	.0617
4	5	.0206
5	1	.0041

Appendix E

Multinomial Reference Distribution of Three-position Grand Totals with 5 Repetitions of 2 Relevant Questions

Score	Ways	Pmf	Cdf	Cdfcontcor	Odds	Oddsll05
-14	153	.0006*	.0011	.0008	1286	9.29
-13	171	.0012	.0023	.0017	587.2	9.04
-12	190	.0024	.0047	.0035	283.9	8.58
-11	210	.0043	.0090	.0069	144.8	7.84
-10	231	.0074	.0162	.0127	77.66	6.8
-9	250	.0119	.0279	.0224	43.64	6.35
-8	267	.0182	.0457	.0376	25.6	5.36
-7	282	.0265	.0711	.0602	15.61	4.22
-6	295	.0366	.1058	.0921	9.86	3.31
-5	306	.0480	.1504	.1348	6.42	2.54
-4	315	.0599	.2053	.1891	4.29	1.89
-3	322	.0712	.2696	.2548	2.92	1.39
-2	327	.0804	.3417	.3304	2.03	1.01
-1	330	.0865	.4194	.4133	1.42	0.72
0	331	.0887	.5000	.5000	1	0.51
1	330	.0865	.5806	.5867	1.42	0.72
2	327	.0804	.6583	.6696	2.03	1.01
3	322	.0712	.7304	.7452	2.92	1.39
4	315	.0599	.7947	.8109	4.29	1.89
5	306	.0480	.8496	.8652	6.42	2.54
6	295	.0366	.8942	.9079	9.86	3.31
7	282	.0265	.9289	.9398	15.61	4.22
8	267	.0182	.9543	.9624	25.6	5.36
9	250	.0119	.9721	.9776	43.64	6.35
10	231	.0074	.9838	.9873	77.66	6.8
11	210	.0043	.9910	.9931	144.8	7.84
12	190	.0024	.9953	.9965	283.9	8.58
13	171	.0012	.9977	.9983	587.2	9.04
14	153	.0006*	.9989	.9992	1286	9.29

* extreme values omitted

Appendix F

Multinomial Reference Distribution of Three-position Grand Totals with 5 Repetitions of 3 Relevant Questions

Score	Ways	*Pmf*	*Cdf*	Cdfcontcor	Odds	Oddsll05
-17	435	.0006*	.0012	.0009	1123	13.95
-16	465	.0010	.0022	.0017	600.6	13.49
-15	496	.0017	.0038	.0030	332.8	12.76
-14	525	.0028	.0066	.0052	190.8	11.7
-13	552	.0044	.0109	.0088	113	10.32
-12	577	.0066	.0175	.0143	69	9.97
-11	600	.0098	.0271	.0225	43.4	8.82
-10	621	.0139	.0406	.0344	28.07	7.35
-9	640	.0190	.0591	.0510	18.62	5.87
-8	657	.0253	.0834	.0733	12.65	4.73
-7	672	.0324	.1142	.1023	8.78	3.77
-6	685	.0401	.1519	.1388	6.21	2.94
-5	696	.0481	.1966	.1831	4.46	2.28
-4	705	.0558	.2480	.2351	3.25	1.76
-3	712	.0626	.3053	.2943	2.4	1.35
-2	717	.0679	.3674	.3593	1.78	1.03
-1	720	.0714	.4328	.4285	1.33	0.78
0	721	.0725	.5000	.5000	1	0.59
1	720	.0714	.5672	.5715	1.33	0.78
2	717	.0679	.6326	.6407	1.78	1.03
3	712	.0626	.6947	.7057	2.4	1.35
4	705	.0558	.7520	.7649	3.25	1.76
5	696	.0481	.8034	.8169	4.46	2.28
6	685	.0401	.8481	.8612	6.21	2.94
7	672	.0324	.8858	.8977	8.78	3.77
8	657	.0253	.9166	.9267	12.65	4.73
9	640	.0190	.9409	.9490	18.62	5.87
10	621	.0139	.9594	.9656	28.07	7.35
11	600	.0098	.9729	.9775	43.4	8.82
12	577	.0066	.9825	.9857	69	9.97
13	552	.0044	.9891	.9912	113	10.32
14	525	.0028	.9934	.9948	190.8	11.7
15	496	.0017	.9962	.9970	332.8	12.76
16	465	.0010	.9978	.9983	600.6	13.49
17	435	.0006*	.9988	.9991	1123	13.95
* extreme values omitted						

Appendix G

Multinomial Reference Distribution of Three-position Grand Totals with 5 Repetitions of 4 Relevant Questions

Score	Ways	Pmf	Cdf	Cdfcontcor	Odds	Oddsll05
-19	900	.0007*	.0016	.0013	782.8	18.12
-18	937	.0011	.0027	.0021	464.3	17.26
-17	972	.0017	.0044	.0035	282.4	16.06
-16	1005	.0026	.0069	.0057	176	14.5
-15	1036	.0038	.0107	.0088	112.3	14.5
-14	1065	.0055	.0161	.0135	73.3	12.21
-13	1092	.0077	.0237	.0200	48.9	10.9
-12	1117	.0105	.0340	.0292	33.3	9.24
-11	1140	.0140	.0476	.0415	23.13	7.57
-10	1161	.0182	.0653	.0576	16.36	6.28
-9	1180	.0230	.0875	.0783	11.76	5.17
-8	1197	.0285	.1147	.1042	8.59	4.15
-7	1212	.0343	.1472	.1358	6.37	3.33
-6	1225	.0403	.1851	.1732	4.77	2.64
-5	1236	.0462	.2282	.2165	3.62	2.09
-4	1245	.0516	.2761	.2654	2.77	1.66
-3	1252	.0563	.3281	.3193	2.13	1.31
-2	1257	.0599	.3835	.3771	1.65	1.03
-1	1260	.0621	.4412	.4378	1.28	0.81
0	1261	.0629	.5000	.5000	1	0.63
1	1260	.0621	.5588	.5622	1.28	0.81
2	1257	.0599	.6165	.6229	1.65	1.03
3	1252	.0563	.6719	.6807	2.13	1.31
4	1245	.0516	.7239	.7346	2.77	1.66
5	1236	.0462	.7718	.7835	3.62	2.09
6	1225	.0403	.8149	.8268	4.77	2.64
7	1212	.0343	.8528	.8642	6.37	3.33
8	1197	.0285	.8853	.8958	8.59	4.15
9	1180	.0230	.9125	.9217	11.76	5.17
10	1161	.0182	.9347	.9424	16.36	6.28
11	1140	.0140	.9524	.9585	23.13	7.57
12	1117	.0105	.9660	.9708	33.3	9.24
13	1092	.0077	.9764	.9800	48.9	10.9
14	1065	.0055	.9839	.9865	73.3	12.21
15	1036	.0038	.9893	.9912	112.3	14.5
16	1005	.0026	.9931	.9944	176	14.5
17	972	.0017	.9956	.9965	282.4	16.06
18	937	.0011	.9973	.9979	464.3	17.26
19	900	.0007*	.9984	.9987	782.8	18.12
* extreme values omitted						

Appendix H

Multinomial Reference Distribution of Three-position Subtotals with 5 Repetitions

Score	Ways	*Pmf*	*Cdf*	Cdf Contcor	Odds	Odds 2rqs	Odds 3rqs	Odds 4rqs	Odds Ll05	Odds2rqll05	Odds3rqll05	Odds4rqll05
-10	21	.0007	.0009*	.0006	1676	40.94	11.88	6.4	4.47	3.24	2.34	1.72
-9	28	.0020	.0029	.0019	514.3	22.68	8.01	4.76	4.37	2.98	1.93	1.42
-8	36	.0050	.0079	.0055	182	13.49	5.67	3.67	4.14	2.52	1.58	1.18
-7	45	.0110	.0188	.0136	72.83	8.53	4.18	2.92	3.7	2.02	1.28	0.99
-6	55	.0214	.0395	.0300	32.38	5.69	3.19	2.39	3.02	1.58	1.06	0.84
-5	66	.0370	.0746	.0597	15.75	3.97	2.51	1.99	2.53	1.26	0.88	0.72
-4	75	.0577	.1277	.1079	8.27	2.88	2.02	1.7	1.98	0.98	0.73	0.62
-3	82	.0810	.1998	.1781	4.61	2.15	1.67	1.47	1.39	0.77	0.61	0.54
-2	87	.1029	.2892	.2704	2.7	1.64	1.39	1.28	0.93	0.6	0.52	0.47
-1	90	.1188	.3912	.3803	1.63	1.28	1.18	1.13	0.6	0.47	0.44	0.41
0	91	.1246	.5000	.5000	1	1	1	1	0.37	0.37	0.37	0.37
1	90	.1188	.6088	.6197	1.63	2.66	4.33	7.05	0.6	0.92	1.32	1.77
2	87	.1029	.7108	.7296	2.7	7.28	19.66	53.05	0.93	1.81	2.81	3.46
3	82	.0810	.8002	.8219	4.61	21.29	98.24	453.3	1.39	2.91	3.88	4.35
4	75	.0577	.8723	.8921	8.27	68.36	565.3	4674	1.98	3.65	4.38	4.49
5	66	.0370	.9254	.9403	15.75	248.2	3911	61610	2.53	4.23	4.49	4.51
6	55	.0214	.9605	.9700	32.38	1048	33950	>99999	3.02	4.44	4.51	4.51
7	45	.0110	.9812	.9865	72.83	5304	>99999	>99999	3.7	4.49	4.51	4.51
8	36	.0050	.9921	.9945	182	33130	>99999	>99999	4.14	4.51	4.51	4.51
9	28	.0020	.9971	.9981	514.3	>100000	>99999	>99999	4.37	4.51	4.51	4.51
10	21	.0007	.9991*	.9994	1676	>100000	>99999	>99999	4.47	4.51	4.51	4.51

* extreme values omitted

Appendix I

Multinomial Reference Distribution for Three-position Grand Totals with 5 Repetitions of 2 Relevant Questions with PLE Sensor

Score	Ways	*Pmf*	*Cdf*	Cdfcontcor	Odds	Oddsll05
-16	2845	.0006*	.0012	.0009	1112	12.4
-15	3136	.0011	.0023	.0017	572.2	11.99
-14	3430	.0019	.0042	.0033	306.5	11.34
-13	3724	.0032	.0074	.0058	170.6	10.36
-12	4015	.0052	.0126	.0101	98.52	9.08
-11	4300	.0080	.0205	.0167	58.91	8.67
-10	4576	.0120	.0322	.0267	36.39	7.55
-9	4840	.0171	.0489	.0414	23.18	6.19
-8	5089	.0235	.0715	.0618	15.18	4.86
-7	5320	.0311	.1011	.0893	10.2	3.96
-6	5530	.0395	.1383	.1248	7.01	3.03
-5	5716	.0485	.1833	.1691	4.92	2.35
-4	5875	.0573	.2360	.2221	3.5	1.79
-3	6004	.0652	.2954	.2833	2.53	1.36
-2	6100	.0714	.3603	.3514	1.85	1.02
-1	6160	.0755	.4292	.4244	1.36	0.76
0	6181	.0769	.5000	.5000	1	0.57
1	6160	.0755	.5708	.5756	1.36	0.76
2	6100	.0714	.6397	.6486	1.85	1.02
3	6004	.0652	.7046	.7167	2.53	1.36
4	5875	.0573	.7640	.7779	3.5	1.79
5	5716	.0485	.8167	.8309	4.92	2.35
6	5530	.0395	.8617	.8752	7.01	3.03
7	5320	.0311	.8989	.9107	10.2	3.96
8	5089	.0235	.9285	.9382	15.18	4.86
9	4840	.0171	.9511	.9586	23.18	6.19
10	4576	.0120	.9678	.9733	36.39	7.55
11	4300	.0080	.9795	.9833	58.91	8.67
12	4015	.0052	.9874	.9900	98.52	9.08
13	3724	.0032	.9926	.9942	170.6	10.36
14	3430	.0019	.9958	.9967	306.5	11.34
15	3136	.0011	.9977	.9983	572.2	11.99
16	2845	.0006*	.9988	.9991	1112	12.4

* extreme values omitted

Appendix J

Multinomial Reference Distribution for Three-position Grand Totals with 5 Repetitions of 3 Relevant Questions with PLE Sensor

Score	Ways	*Pmf*	*Cdf*	Cdfcontcor	Odds	Oddsll05
-19	12100	.0007*	.0016	.0013	782.8	18.12
-18	12734	.0011	.0027	.0021	464.3	17.26
-17	13360	.0017	.0044	.0035	282.4	16.06
-16	13975	.0026	.0069	.0057	176	14.5
-15	14576	.0038	.0107	.0088	112.3	14.5
-14	15160	.0055	.0161	.0135	73.3	12.21
-13	15724	.0077	.0237	.0200	48.9	10.9
-12	16265	.0105	.0340	.0292	33.3	9.24
-11	16780	.0140	.0476	.0415	23.13	7.57
-10	17266	.0182	.0653	.0576	16.36	6.28
-9	17720	.0230	.0875	.0783	11.76	5.17
-8	18139	.0285	.1147	.1042	8.59	4.15
-7	18520	.0343	.1472	.1358	6.37	3.33
-6	18860	.0403	.1851	.1732	4.77	2.64
-5	19156	.0462	.2282	.2165	3.62	2.09
-4	19405	.0516	.2761	.2654	2.77	1.66
-3	19604	.0563	.3281	.3193	2.13	1.31
-2	19750	.0599	.3835	.3771	1.65	1.03
-1	19840	.0621	.4412	.4378	1.28	0.81
0	19871	.0629	.5000	.5000	1	0.63
1	19840	.0621	.5588	.5622	1.28	0.81
2	19750	.0599	.6165	.6229	1.65	1.03
3	19604	.0563	.6719	.6807	2.13	1.31
4	19405	.0516	.7239	.7346	2.77	1.66
5	19156	.0462	.7718	.7835	3.62	2.09
6	18860	.0403	.8149	.8268	4.77	2.64
7	18520	.0343	.8528	.8642	6.37	3.33
8	18139	.0285	.8853	.8958	8.59	4.15
9	17720	.0230	.9125	.9217	11.76	5.17
10	17266	.0182	.9347	.9424	16.36	6.28
11	16780	.0140	.9524	.9585	23.13	7.57
12	16265	.0105	.9660	.9708	33.3	9.24
13	15724	.0077	.9764	.9800	48.9	10.9
14	15160	.0055	.9839	.9865	73.3	12.21
15	14576	.0038	.9893	.9912	112.3	14.5
16	13975	.0026	.9931	.9944	176	14.5
17	13360	.0017	.9956	.9965	282.4	16.06
18	12734	.0011	.9973	.9979	464.3	17.26
19	12100	.0007*	.9984	.9987	782.8	18.12
* extreme values omitted						

Appendix K

Multinomial Reference Distribution for Three-position Grand Totals with 5 Repetitions of 4 Relevant Questions with PLE Sensor

Score	Ways	*Pmf*	*Cdf*	Cdfcontcor	Odds	Oddsll05
-22	31430	.0006*	.0015	.0013	796.1	23.78
-21	32500	.0009	.0024	.0020	506.7	22.59
-20	33551	.0013	.0037	.0030	328.5	21.02
-19	34580	.0018	.0055	.0046	216.8	19.08
-18	35584	.0026	.0081	.0068	145.6	19.31
-17	36560	.0036	.0117	.0100	99.5	16.57
-16	37505	.0050	.0166	.0143	69.12	15.16
-15	38416	.0067	.0232	.0201	48.79	13.24
-14	39290	.0087	.0318	.0278	34.97	11.18
-13	40124	.0113	.0429	.0378	25.43	9.56
-12	40915	.0142	.0568	.0506	18.75	7.94
-11	41660	.0176	.0739	.0666	14.01	6.49
-10	42356	.0215	.0947	.0863	10.59	5.33
-9	43000	.0256	.1193	.1099	8.1	4.38
-8	43589	.0300	.1480	.1379	6.25	3.59
-7	44120	.0346	.1807	.1702	4.87	2.92
-6	44590	.0390	.2175	.2071	3.83	2.38
-5	44996	.0432	.2580	.2482	3.03	1.93
-4	45335	.0470	.3019	.2933	2.41	1.57
-3	45604	.0501	.3488	.3417	1.93	1.28
-2	45800	.0525	.3979	.3929	1.55	1.04
-1	45920	.0540	.4485	.4460	1.24	0.84
0	45961	.0545	.5000	.5000	1	0.68
1	45920	.0540	.5515	.5540	1.24	0.84
2	45800	.0525	.6021	.6071	1.55	1.04
3	45604	.0501	.6512	.6583	1.93	1.28
4	45335	.0470	.6981	.7067	2.41	1.57
5	44996	.0432	.7420	.7518	3.03	1.93
6	44590	.0390	.7825	.7929	3.83	2.38
7	44120	.0346	.8193	.8298	4.87	2.92
8	43589	.0300	.8520	.8621	6.25	3.59
9	43000	.0256	.8807	.8901	8.1	4.38
10	42356	.0215	.9053	.9137	10.59	5.33
11	41660	.0176	.9261	.9334	14.01	6.49
12	40915	.0142	.9432	.9494	18.75	7.94
13	40124	.0113	.9571	.9622	25.43	9.56
14	39290	.0087	.9682	.9722	34.97	11.18
15	38416	.0067	.9768	.9799	48.79	13.24
16	37505	.0050	.9834	.9857	69.12	15.16
17	36560	.0036	.9883	.9900	99.5	16.57
18	35584	.0026	.9919	.9932	145.6	19.31
19	34580	.0018	.9945	.9954	216.8	19.08
20	33551	.0013	.9963	.9970	328.5	21.02
21	32500	.0009	.9976	.9980	506.7	22.59
22	31430	.0006*	.9985	.9987	796.1	23.78

* extreme values omitted

Appendix L

Multinomial Reference Distribution of Three-position Subtotals with 5 Repetitions with PLE Sensor

Score	Ways	*Pmf*	*Cdf*	Cdf Contcor	Odds	Odds 2rqs	Odds 3rqs	Odds 4rqs	Odds Ll05	Odds2rqll05	Odds3rqll05	Odds4rqll05
-12	165	.0004	.0006*	.0004	2517	50.17	13.6	7.08	6.13	4.35	2.99	2.13
-11	220	.0010	.0016	.0011	888.3	29.8	9.61	5.46	6.03	4.11	2.54	1.85
-10	286	.0024	.0041	.0029	346.6	18.62	7.03	4.32	5.82	3.58	2.11	1.56
-9	360	.0052	.0092	.0067	148.1	12.17	5.29	3.49	5.41	2.96	1.8	1.32
-8	439	.0100	.0191	.0144	68.67	8.29	4.1	2.88	4.73	2.39	1.49	1.15
-7	520	.0177	.0364	.0284	34.27	5.85	3.25	2.42	4.35	1.9	1.26	0.99
-6	600	.0289	.0640	.0520	18.25	4.27	2.63	2.07	3.54	1.54	1.07	0.87
-5	676	.0434	.1047	.0886	10.28	3.21	2.18	1.79	2.65	1.25	0.9	0.76
-4	745	.0605	.1601	.1412	6.08	2.47	1.83	1.57	1.96	1.01	0.77	0.67
-3	804	.0781	.2299	.2110	3.74	1.93	1.55	1.39	1.4	0.81	0.66	0.6
-2	850	.0937	.3123	.2969	2.37	1.54	1.33	1.24	0.97	0.66	0.57	0.54
-1	880	.1044	.4037	.3951	1.53	1.24	1.15	1.11	0.66	0.53	0.5	0.48
0	891	.1082	.5000	.5000	1	1	1	1	0.43	0.43	0.43	0.43
1	880	.1044	.5963	.6049	1.53	2.34	3.59	5.5	0.66	0.96	1.36	1.86
2	850	.0937	.6877	.7031	2.37	5.61	13.28	31.46	0.97	1.89	2.95	4.21
3	804	.0781	.7701	.7890	3.74	13.98	52.25	195.3	1.4	3.04	4.41	5.58
4	745	.0605	.8399	.8588	6.08	36.97	224.8	1367	1.96	4.47	5.65	6.08
5	676	.0434	.8953	.9114	10.28	105.7	1087	11180	2.65	5.15	6.06	6.16
6	600	.0289	.9360	.9480	18.25	333	6076	>99999	3.54	5.81	6.16	6.18
7	520	.0177	.9637	.9716	34.27	1174	40240	>99999	4.35	6.07	6.17	6.18
8	439	.0100	.9809	.9856	68.67	4716	>99999	>99999	4.73	6.15	6.18	6.18
9	360	.0052	.9908	.9933	148.1	21930	>99999	>99999	5.41	6.17	6.18	6.18
10	286	.0024	.9959	.9971	346.6	>99999	>99999	>99999	5.82	6.18	6.18	6.18
11	220	.0010	.9984	.9989	888.3	>99999	>99999	>99999	6.03	6.18	6.18	6.18
12	165	.0004	.9994*	.9996	2517	>99999	>99999	>99999	6.13	6.18	6.18	6.18

* extreme values omitted

Erratum Notice

It has come to our attention a typographical error exists on page 212 of 'Polygraph & Forensic Credibility Assessment' *Polygraph*, 2017, 46 (2). At the top of the page, parasympathetic should be sympathetic.

The correct document may be downloaded from the below link, please log in to your APA account before clicking on the link to view the article.

https://apoa.memberclicks.net/assets/docs/APA-Journal.Articles/2017/journal%20462%20preview%203.pdf

We apologize for any inconvenience this may have caused.

AMERICAN POLYGRAPH ASSOCIATION

DEDICATED TO TRUTH