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

VOLUME 48 **2019** **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

Subscription information: *Polygraph* is published semi-annually by the American Polygraph Association. Editorial Address is Editor@polygraph.org. Subscription rates for 2019: One year $150.00 (Domestic). 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 Psycho-physiological 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.

Analyzing Iacono's Thought Experiment About Polygraph Field Studies: Reason or Fantasy?

Charles R. Honts

Boise State University

and

Steven Thurber

Minnesota Department of Human Services

Abstract

We review and analyze a thought experiment first published in Iacono (1991) and reintroduced in Iacono and Ben-Shakhar (2019). The Iacono Thought Experiment (ITE) appears to have used backtracking methods to generate a series of assumptions and preconditions which would make it possible to have a polygraph test with chance accuracy that produces a confession-criterion field study with high accuracy. From this thought experiment, Iacono promulgated a hypothesis that all polygraph confession criterion studies produce exaggeratedly high estimates of accuracy. Our analysis of the assumptions and preconditions of the ITE found them to be unrepresentative and highly unlikely to be met in real world settings. We used a converging evidence approach that applied meta-analytic results, field studies that did not use a confession criterion, and data from wrongful conviction cases that involved polygraph examinations to test the Iacono hypothesis. We found strong falsification evidence to the Iacono hypothesis and conclude that it should be abandoned as a meaning description of field polygraph research.

Keywords: Ocular-motor, deception detection, eye tracking, reading

Analyzing Iacono's Thought Experiment About Polygraph Field Studies: Reason or Fantasy?

Polygraph tests represent an important and widespread application of a psychological test in law enforcement, national security, and employment around the world. Internationally, the American Polygraph Association shows members from 62 countries. Zhang (2011) estimated that there were as many as 8000 polygraph examiners operating in China alone. Despite the ubiquitous nature of polygraph testing, it has received relatively little attention in academic psychology and often, that attention has been in the form of negative commentary.

The most commonly used - and criticized - polygraph test is the Comparison Question Test (CQT). The CQT comes in sev-

Author Note

Correspondence should be addressed to Charles R. Honts, Ph. D., Department of Psychological Science, Boise State University, 1910 University Drive MS-1715, Boise ID 83725-1715. The authors would like to thank Adela Stephanescu for her help in editing the completed manuscript.

eral variants, but in all cases, it monitors the subject's autonomic physiology (usually, respiration, electrodermal activity, relative blood pressure, and often peripheral vasomotor activity) while the subject answers a series of questions. The question series contains two categories of critical questions (usually two or three of each). Relevant questions directly address the matter under investigation. Comparison questions are designed and presented in such a way that every subject lies, or is assumed to lie, in their response to them during the test. The subject's physiological responses are expected to show an interaction, so that subjects who are deceptive to the relevant questions show larger physiological responses to relevant questions as compared to their comparison questions. Subjects who are being truthful to the relevant questions are expected to show the opposite pattern, with physiological responses to comparison questions being larger than those to relevant questions.

There are a number of reviews of CQT research, typified by but not limited to the following examples: Raskin, Honts, and Kircher (1997), Iacono & Lykken (1997), The National Research Council (NRC, 2003), Honts (2004), Vrij (2008); American Polygraph Association (2011); Raskin, Honts, and Kircher (2014), and Iacono and Ben-Shakhar (2018). There is variation across the reviews, but nevertheless they generally produced overall accuracy estimates over 80%.

However, all of those reviews can be criticized for selective study choices and a lack of meta-analytic scrutiny. None attempted to test moderator variables, although they sometimes reached conclusions that hypothesized or even assumed powerful moderator effects. The NRC (2003) report was particularly egregious in that regard. NRC found objectively high discrimination estimating the area under the Receiver Operating Characteristic (ROC) curve (*AUC*) at 0.89. The use of ROC analysis and *AUC* as an effect size has been criticized as an inappropriate application of a technology developed to examine the performance of signal detection with technology (specifically an operator's ability to view RADAR screens and distinguish enemy ships, friendly ships, and noise; Tape, 2019) to psychology in general (Balakrishnan, 1999), and to polygraph testing in particular (Honts & Schweinle, 2009).

Nevertheless, the NRC used *AUC* as their index of effect size, but what does AUC actually mean and what does an *AUC* value of 0.89 imply about polygraph performance? The value of *AUC* can range from 0.50, which represents chance performance, to 1.00 which represents perfect performance (100% accuracy; Tape, 2019). Tape (2019) qualitatively characterizes *AUC* values between 0.80 and 0.90 as indicating a good discriminator and *AUC* values above 0.90 as excellent. Tables (Rice & Harris, 2005) and software (DeCoster, 2012) to convert between *AUC* and other measures of effect size are readily available. Reference to those tables and software show that an *AUC* value of 0.89 corresponded to a *Cohen's d* value of 1.74 and an r_{pb} of 0.66. Cohen (1969, 1988, 1992) described large effects in psychology as those with d values above 0.80 (corresponding $r_{pb} > 0.49$). Cohen famously said that, in applied psychology, effect sizes of d = 0.8 are "about as high as they come" (Cohen, 1988, p. 81). Thus, the *AUC* effect size reported by NRC (2003) indicates extremely high performance for the CQT as compared to other psychological tests and measures.

In spite of powerful empirical evidence of the usefulness of the CQT as a discriminator of truth and deception, the NRC discounted those findings, saying the research methods were substandard. To the present authors this seems to be an arrogant conclusion as the NRC substituted their judgment about the qualities of research published in first tier peer-reviewed journals of psychological science. Such a position is insulting to the editors of those first-tier journals and the working scientists who peer-review for them. The NRC's opinion is all the worse for the fact that none of the members of the NRC committee who wrote the report had ever published a study on deception detection.

Additionally, the NRC and others were notably critical of the use of experimental (laboratory) studies for assessing CQT validity. Iacono & Lykken (1997) completely dismiss the experimental research on the CQT, arguing that the real-world motivational contexts could not be modeled experimentally and therefore laboratory results were qualitatively different from those in real cases. However, the conclusions of the NRC and Iacono & Lykken (1997) about the generalizability of the

research methods in archival peer-reviewed journals of psychological science and the generalizability of experimental CQT research should all be viewed as opinion and not as fact, as none of those opinions were or are data-based.

The history of academic disagreement over the accuracy (criterion validity) of the CQT is long and has at times been polemic. Those disagreements are typified by, but are not limited to, published exchanges between researchers from the University of Utah and the University of Minnesota, beginning in 1978 in the journals *Psychophysiology* (Raskin and Hare, 1978; Lykken, 1978; Raskin, 1978) and *Psychological Bulletin* (Lykken, 1979; Raskin and Podlesny, 1979). Direct exchanges in the literature between these groups continued until 2002 (Honts, Raskin, and Kircher, 2002; Iacono and Lykken, 2002). Those disagreements were argued at a number of different levels on various topics. Throughout the disagreements, the Minnesota group radically rejected the notion that deception detection could be validly modeled in the laboratory and held that the results of laboratory studies were not useful for estimating criterion validity in field applications because they lacked external validity (generalizability). The Minnesota group holds that position until today despite the general rejection of such criticisms across the entirety of Psychological Science and specifically for deception detection. Hartwig and Bond (2014) provided a general discussion about generalizability of laboratory studies and provided a specific empirical rejection of differences between experimental and field settings, within a meta-analysis of the interpersonal deception detection research literature.

The scientific issues surrounding the contrast of experimental and field settings for research in interpersonal deception detection are nearly identical to those with the CQT. Recently, Honts and Thurber (2019) reported a comprehensive meta-analysis of the CQT that followed the analytic approach of Hartwig and Bond (2019). Honts and Thurber (2019) reported no statistically detectable effects for moderators of motivation, subject population or setting (experiment vs. field) in their comprehensive meta-analysis of the CQT.

The Minnesota group was initially supportive of field studies that fit their criterion for useful field studies. However, starting in the 1980s, field studies were published that produced high levels of accuracy with the CQT (Honts & Raskin, 1988; Raskin, Kircher, Honts, & Horowitz, 1988). Those studies were specifically designed to meet the Minnesota group's criteria. Subsequently, the Minnesota group rejected all field studies and took the radical position that valid research on the CQT could not be conducted. One keystone of that position was a thought experiment first reported by Iacono (1991) and then with some modification reintroduced in Iacono and Ben-Shakhar (2018)[1]. The Iacono (1991) thought experiment was originally presented as follows:

> Suppose that 800 crimes are being investigated using a polygraph technique that operates with exactly chance accuracy; i.e. half of both the guilty and innocent suspects will fail and half will pass. Because the polygraph is often used in crimes for which there are multiple suspects, let us assume, without loss of generality, that we are dealing with 800 two-suspect crimes, and that for each, one suspect is guilty and the other innocent. Let us assume further that (1) the guilty suspect is tested first 50% of the time, (2) the second suspect will not be tested if the result of the first test indicates deception, (3) neither innocent suspects nor those guilty suspects who pass the test will confess, and (4) 20% of the guilty who fail the test and are subsequently interrogated confess. (Iacono, 1991, pp. 202-203).

[1]In Iacono and Ben-Shakhar provided two simplified versions of the Iacono Thought Experiment with single subjects and with paired subjects. Assumptions 3, 4, and 5 do not apply to either the single subject or the paired tests as all subjects are tested regardless of the outcomes. Confession rates are not specified for either analysis thus Assumption 7 is not specific. The other assumptions are either explicit or implied in the latter version of the Iacono Thought Experiment.

Thought experiments are well known in philosophy and science. Thought experiments can be defined as "devices of the imagination used to investigate the nature of things" (Brown, 2014). One of the most famous scientific thought experiments was Galileo's reasoning that two objects of different weight must fall at the same speed. Galileo's thought experiment is easily validated from observations, such as the conclusive demonstration by Neil Armstrong on the moon when he dropped a feather and a hammer simultaneously and they landed on the surface at the same time (Pigllucci, 2006). Pigllucci further notes that thought experiments can also be wrong and be falsified by data. Had Galileo's thought experiment been invalidated by data, it would have been lost to history and forgotten.

Thought experiments can take on a number of forms or types. While a discussion of the multiple types of thought experiments is beyond the scope of this paper, it is worth noting that Iacono's thought experiment appears to be a type known as Backcasting (Robinson, 1982). In Backcasting one imagines a desired or possible state of the world and then reasons backward from that end-state to the necessary precursors. By definition, such logic necessarily does not provide a description of reality, it only provides a chain of precursors that might produce the desired end-state. Such thought experiments, like all thought experiments, are useful in the real world only to the extent that they can be tested and validated or falsified with data. We begin our analysis of the Iacono Thought Experiment (ITE) by defining the hypothetical precursors that he either invented or selected to reach the desired end-state where polygraph tests with chance accuracy could produce a field study with high accuracy rates.

Elucidation and Analysis of the Hypothetical Preconditions and Assumption of the Iacono Thought Experiment

Explicit Assumptions of the Iacono (1991) Thought Experiment.

Iacono (1991) makes a number of explicit assumptions that were used to create a possible path to the desired end state.

1.Eight hundred subjects are tested where 400 are Innocent and 400 are Guilty.

2.The polygraph preforms exactly at chance accuracy of 50% correct, 50% incorrect, and no inconclusive outcomes. This assumption is part of the overall desired end-state where a chance polygraph test could produce high accuracy outcomes. All of the other assumptions also serve the establishment of that end-state.

3.Each crime has only two suspects. (Iacono makes this assumption and states that it is made "without loss of generality" (p. 203).

4.The Guilty suspect is tested first in half of the cases

5.If the first suspect fails the polygraph test, the second suspect will not be tested.

6.Neither innocent nor guilty suspects who pass the test will confess.

7.Only 20% of the Guilty suspects who fail and are interrogated will confess.

Implicit Assumptions of the Iacono Thought Experiment. The following implicit assumptions are also necessary for the mathematics and logic of the Iacono Thought Experiment to reach the desired end-state.

8.The polygraph is the only source of information about who is guilty in a criminal case.

9.Guilty people only confess after polygraph examinations.

Analysis of the Explicit Assumptions of the Iacono Thought Experiment

Experiment

Assumption 1 is that the base rate of guilty to innocent subjects is equal. The base rate of guilt in a criminal case will vary greatly depending upon when the polygraph is used. If it is used early in an investigation, there are likely to be far more innocent than guilty sub-

jects; if it is used very late in an investigation, there may be many more guilty than innocent subjects. The assumption of equal base rates is acceptable for a thought experiment as long as one recognizes that variations in the base rate could dramatically alter the end state results and that a base rate of 50% will be unusual in actual practice.

Assumption 2 is that the polygraph performs exactly as a coin flip. This assumption is made as a premise of the thought experiment and it is a necessary component of the desired end state. However, this premise is without empirical support in the real world. To our knowledge, there are no studies that show any version of the CQT to perform at chance levels.

Assumption 3 is that each case has only two suspects. This premise simplifies the mathematics necessary to achieve the desired end state of the Iacono thought experiment, but it is a premise that is rarely met in the real world, and is not at all representative of the field at the time Iacono (1991) was written (for example, Honts & Raskin, 1988, Raskin, Kircher, Honts, & Horowitz, 1988 all contain many single and multiple suspect cases as do the more recent field studies). Iacono's assertion that this assumption is made without a loss of generality (for the Backcasting thought experiment) is clearly not supported by data.

Assumption 4 is that a guilty suspect is tested first in half of the cases. This assumption is tenable only if there are only two suspects and that the examiner has no reason to test one or the other suspect first. It is a convenient assumption for the thought experiment, but it is unlikely to be widely representative of field polygraph testing.

Assumption 5 states that if the first person is tested and fails the second suspect will not be tested. However, this is not the case in real investigations. If the first subject is tested and fails but does not confess, then the remaining suspect or suspects will likely be tested to assess their involvement in the crime. The likely logic of investigators would be that the additional suspect(s) would not be suspects unless there was a reason to suspect them, and they may well be involved. In our experience it is, in fact, common practice to test all suspects in a case during an investigation.

Assumption 6 states that neither innocent suspects nor guilty suspects who pass the test will confess. This is manifestly not true. Under certain circumstances, such as a wrongfully failed or deliberately misrepresented polygraph test result, innocent suspects will confess to crimes they did not commit. The White Paper of the American Psychology Law Society (Kassin, Drizin, Grisso, Gudjonsson, Leo, & Redlich, 2010) specifically notes that wrongfully failed or willfully misrepresented polygraph outcomes are a powerful false evidence ploy that puts the actually innocent at increased risk of false confession. Moreover, Bonpasse (2013), provides examples and discussion of actual cases where incorrect or misrepresented polygraph outcomes have contributed to miscarriages of justice through their role in eliciting false confessions. Assumption 6 is also incorrect for guilty suspects who pass polygraphs, as it ignores the fact that investigations rarely stop just because a polygraph has been passed. If the subsequent investigation continues and additional information is obtained, then the suspect will likely be interviewed a second time, despite the passed polygraph, and may provide a confession then or confess later as part of a plea bargain. At least one such case was included in Honts & Raskin (1988).

Assumption 7 this assumption states that only 20% of the guilty suspects who are interrogated will confess. The choice of 20% is arbitrary and has no empirical basis. The actual confession rate will depend upon the situation in which the tests were conducted. Polygraph tests conducted for defense attorneys, or by the police on subjects who have defense counsel, are unlikely to be followed up with interrogations regardless of the polygraph outcome. On the other hand, the U. S. Department of Defense (2002) has reported data indicating that in one polygraph program, more than 90% of the failed polygraph examinations resulted in relevant admissions. Clearly, the rate chosen for this assumption will have a major impact on the resultant outcomes of the ITE. Moreover, since all confession rate values are situationally specific, it is non-sensical to provide a single value for central tendency as such a value would be meaningless to any

specific applied setting.

Analysis of the Implicit Assumptions of the Iacono Thought Experiment

Assumption 8 asserts that the polygraph is the only source of information about who is guilty in a criminal investigation. This ignores that fact that individuals can confess in contexts other than polygraph examinations, or that other incontrovertible evidence of guilt or innocence may be obtained independent of the polygraph test. This was examined explicitly in Honts (1996) and no differences in numerical scores or accuracy rates were found between confession confirmed and evidence confirmed cases for guilty or innocent subjects. Interestingly, in Honts (1996) none of the innocent subjects were confirmed by a confession obtained in the context of a polygraph examination. Assumption 8 is preposterous on its face, but the ITE cannot work without it.

Assumption 9 is that guilty people only confess following polygraph tests. As covered in our discussion of Assumption 6, we noted that inaccurate and misrepresented polygraph tests can result in false confessions from innocent subjects. We also noted that even guilty subjects who pass polygraph tests will sometimes confess later, when faced with new or overwhelming evidence. Moreover, guilty suspects, and occasionally innocent subjects, will confess as part of a plea bargain. Thus, it is obviously true that guilty people who fail a polygraph test but either are not interrogated, or initially resist an interrogation, may confess later. At least three field studies have explicitly taken this into account and looked for confirming information in an exhaustive sample of cases within a particular period of time, and used all of the information available not only in the polygraph examination file, but in the complete police record of the case (Honts, 1996; Patrick & Iacono, 1991; Raskin et al., 2019).

Summary of the Analysis of Preconditions and Assumptions

Our analysis shows that the assumptions of the Iacono thought experiment were generally chosen without reference to data or professional practice, in the service of developing what became a highly improbable set of preconditions and assumptions leading to a specific solution showing that a polygraph test with chance accuracy could produce a field study with high accuracy rates. The Iacono thought experiment was then transmogrified into a normative statement that all field studies of the CQT were, are, and forever will be unreliable and overestimate actually accuracy. We do not believe that this normative conclusion is justified unless it can withstand empirical examination and falsification. For the remainder of this paper we will refer to the hypothesis derived from the Iacono thought experiment, that the CQT is no more accurate than chance and that all confession criterion studies are biased to dramatically overestimate the accuracy of the CQT as the Iacono Thought Experiment Hypothesis (ITEH).

Data That Could Falsify the Iacono Thought Experiment Hypothesis (ITEH)

Just like Galileo's thought experiment concerning falling objects, the ITEH survives the test of science based upon a lack of falsification data in the scientific research. This leads to the question of what data would falsify the ITEH? The remainder of this paper addresses several sources of converging data that do, in fact, lead to the conclusion that the ITEH is false.

Convergence of Experimental and Field Data Without Detectable Moderator Effects

Recent studies summarized by Hartwig and Bond (2014) have indicated strong convergence between experimental and field studies in psychological science and interpersonal deception detection. Hartwig and Bond explicitly rejected the notion that experiments and field research on interpersonal deception detection produced significantly different results. If the ITEH were true, then polygraph testing would have to be qualitatively different underlying mechanisms from interpersonal deception. Under such circumstances we would expect that laboratory studies of the CQT would produce dramatically lower ac-

curacies than the (according to the ITEH) exaggeratedly high accuracies produced by the supposedly unavoidable effects of the ITEH on field studies of the CQT. Existing reviews simply do not reveal dramatically more accurate results in field than in the laboratory (NRC, 2003; Honts & Thurber, 2019).

Lack of Differences in Accuracy Between Field Studies that Rely on the Confession Criteria and Those That Do Not.

Since the ITEH is critically bound to the use of confessions as a criterion of confirmation of Guilt and Innocence, and the ITEH predicts that the confession criterion critically biases field studies to show high accuracy, we should expect that field studies that included or used other methods of confirmation would produce accuracies that approach chance levels of accuracy. Empirically, this is simply not the case. Honts (1996) directly tested this hypothesis, rating strength of confirmation on a scale that ranged from confessions with the generation of new evidence at one end of the scale to no confirmation at the other. Honts (1996) tested that scale against decision accuracy and against numerical scores. In direct opposition to the predictions of the ITEH, Honts found no effects for the level of confirmation. That is, confession confirmed cases did not have higher accuracy levels than cases that were confirmed by methods other than confession (physical evidence and/or witness statements).

Similarly, there are two field studies that use paired testing and mathematics to estimate accuracy (Ginton, 2013; Mao, Liang, & Hu, 2015). This paired testing approach, while not without problems (Iacono & Ben-Shakhar, 2018), is not dependent upon confessions and so is outside the scope of the ITEH. Estimated accuracy rates from the paired subjects studies converge with data from both laboratory and field studies and thus provide support for both.

Lack of Concurrence Between Wrongful Convictions and Failed CQT Polygraph Tests.

If the ITEH is correct that CQT polygraphs are no more accurate than chance, we would expect that, on average, half of the innocent subjects tested in criminal justice settings would produce false positive errors. In the criminal justice settings, innocent subjects who failed the polygraph would be exposed to interrogation and thus put at risk of false confession. Under such circumstances, we would expect there to be a relatively large number of false positive outcomes among the ranks of the wrongfully convicted. Bonpasse (2013) reviewed the case files of the National Registry of Exonerations[2] which was founded in 1989 as a joint project of the University of Michigan and Northwestern University Law School. Bonpasse reported finding 215 exoneration cases where polygraph tests were involved. Of those 215 cases only 23 (10.7%) contained information that an Innocent subject had been tested before trial and had failed the polygraph. However, there were 44 (20.5%) Innocent subjects who had been tested with the polygraph before trial, produced truthful outcomes, but those favorable outcomes did not help them avoid wrongful conviction. Although the ITEH predicts that false positive errors should be common among the wrongfully convicted, they occurred at only half the rate of true negative outcomes. Bonpasse also reported that across all testing, before and after trial and including tests of the immediate suspect and others (co-defendants and witnesses), 135 (62.9%) of the polygraph test outcome were favorable to the wrongfully convicted person while only 31 (14.4%) produced unfavorable outcomes. Data from the wrongfully convicted strongly contradicts and thus falsifies ITEH.

Discussion

To our knowledge, there is not a single study of the CQT, either laboratory or field, that produced chance accuracy rates. While there is a substantial amount of variability between studies of the CQT, no review has found

[2]http://www.law.umich.edu/special/exoneration/Pages/about.aspx

that laboratory studies are dramatically less accurate than field studies (NRC, 2003; Honts & Thurber, 2019). Thus, the ITEH completely lacks empirical substantiation. Moreover, data from the Honts & Thurber (2019) meta-analysis, field studies that do not use the confession criterion, and the wrongfully convicted all provide evidence that the ITEH is false. The results of the ITE are therefore seen as a failed thought experiment that is completely without empirical support, and which should be relegated to the trash heap of history's failed ideas.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph, 40,* 194-305.

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1189–1206.

Bonpasse, M. (2013). Polygraph and 215 wrongful conviction exonerations. *Polygraph, 42,* 112-127.

Brown, J. R. (2014). Thought experiments. *Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/entries/thought-experiment/

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Edition). Hillsdale NJ:Erlbaum.

Cohen, J., (1992). *A power primer. Psychological Bulletin*, 122, 155-159.

DeCoster, J. (2012). Converting effect sizes 2012-06-19-4.xls Retrieved from http://www.stat-help.com/spreadsheets.html

Ginton, A. (2013). A non-standard method for estimating accuracy of lie detection techniques demonstrated on a self-validating set of field polygraph examinations. Psychology, Crime, & Law, 19, 577-594. DOI: http://dx.doi.org/10.1080/1068316X.2012.656118

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology,* 28, 661-676.

Honts, C. R. (1996). Criterion development and validity of the control question test in field application. *The Journal of General Psychology, 123,* 309-324.

Honts, C. R. (2004). The psychophysiological detection of deception, in P. Granhag and L. Strömwall (Eds.) *Detection of deception in forensic contexts.* London: Cambridge University Press 103-123.

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

Honts, C. R., & Schweinle, W., (2009). Information gain of psychophysiological detection of deception in forensic and screening settings. *Applied Psychophysiology and Biofeedback, 34,* 161-172. (Available online July 2009)

Honts, C. R., & Thurber, S. (2019, March). A Comprehensive Meta-Analysis of the Comparison *Question Polygraph Test.* Paper presented at the annual meeting of the American Psychology Law Society, Portland, Oregon.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2002). The scientific status of research on polygraph techniques: The case for polygraph tests. In, D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.) Science in the Law: Social and Behavioral Sciences Issue, American Casebook Series (pp. 598-634). West Group: St. Paul, Minnesota.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? In J. R. Jennings, P.K.

Ackles & M. G. H. Coles (Eds.) *Advances in psychophysiology* (pp. 201-207). London, UK: Jessica Kingsley Publishers.

Iacono, W. G., & Ben-Shakhar, G. (2019). Current Status of forensic lie detection with the comparison question test: An Update of the 2003 National Academy of Sciences report on polygraph testing. *Law and Human Behavior, 43*, 86-98.

Iacono, W. G., & Lykken, D. T. (1997). The scientific status of research on polygraph techniques: The case against polygraph tests. In, D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.) *Science in the Law: Social and Behavioral Sciences Issue, American Casebook Series* (pp. 582-618-). West Group: St. Paul, Minnesota.

Iacono, W. G., & Lykken, D. T. (2002). The scientific status of research on polygraph techniques: The case against polygraph tests. In, D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.) *Science in the Law: Social and Behavioral Sciences Issue, American Casebook Series* (pp. 634-688). West Group: St. Paul, Minnesota.

Kassin, S. M., Drizin, S. A., Grisso, T., Gudjonsson, G. H., Leo, R. A., & Redlich, A. D. (2010). Police-induced confessions: Risk factors and recommendations. *Law and Human Behavior,* 34, 3-38.

Lykken, D. T. (1978). The psychopath and the lie detector. *Psychophysiology*, 15, 137–142. http://dx.doi.org/10.1111/j.1469-8986.1978.tb01349.x

Lykken, D. T. (1979). The detection of deception. *Psychological Bulletin,* 86, 47–53. http://dx.doi.org/10.1037/0033-2909.86.1.47

Mao, Y., Liang, Y., & Hu, Z. (2015). Accuracy rate of lie-detection in China: Estimate the validity of CQT on field cases. *Physiology & Behavior,* 140, 104-110.

National Research Council (2003). *The Polygraph and Lie Detection.* Washington, DC: The National Academies Press.

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

Piglucci, M. (2006). What is a thought experiment, anyhow? *Philosophy Now: A magazine of Ideas, 58,* Retrieved from

https://philosophynow.org/issues/58/What_is_a_Thought_Experiment_Anyhow

Raskin, D. C. (1978). Scientific assessment of the accuracy of deception of detection: A reply to Lykken. *Psychophysiology, 15,* 143-147.

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

Raskin, D. C., & Podlesny, J. A. (1979). Truth and deception: A reply to Lykken. P*sychological Bulletin, 86,* 54–59. http://dx.doi.org/10.1037/0033-2909.86.1.54

Raskin, D. C., Honts, C. R., & Kircher, J. C. (1997). The scientific status of research on polygraph techniques: The case for polygraph tests. Chapter in, D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.) Modern scientific evidence: The law and science of expert testimony (pp. 565-582).

Raskin, D. C., Honts, C. R., & Kircher, J. C. (2014). *Credibility assessment: Scientific research and applications.* Oxford, UK: Academic Press. ISBN: 978-0-12-394433-7 (ebook version available online 17 December 2013).

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (2019). A study of the validity of polygraph examinations in criminal investigations. Final report to the National Institute of Justice, Grant Number 85-IJ-CX-0400. *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice, 48,* 10-39.

Rice, M. E. and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior,* 29, 615-620.

Robinson, J. B. (1982) Energy backcasting: A proposed method of policy analysis. *Energy Policy*, December.

Tape, T. G. (2019). Interpreting Diagnostic Tests. University of Nebraska Medical Center. Retrieved from, http://gim.unmc.edu/dxtests/Default.htm also referencing subordinate web pages on the same topic at this URL.

U. S. Department of Defense (2002). *Department of Defense Polygraph Program Annual Report to Congress, Fiscal Year 2002.* Office of the Assistant Secretary of Defense (Command, Control, Communications, and Intelligence.

Vrij, A. (2008). Detecting Lies and Deceit: Pitfalls and Opportunities, Second Edition. Chichester, UK: Wiley.

Zhang, X. (2011). The evolution of polygraph testing in the People's Republic of China. *Polygraph,* 40, 181-193.

Structural Optimization of Respiration, EDA and Cardio Activity Using a Genetic AI

Raymond Nelson

Abstract

This project involved the use of a balanced sample of n=36 field polygraph exams and a simple genetic algorithm to compute a weighting function for polygraph signals that would optimize the classification of deception and truth-telling. A genetic algorithm is a simple form of machine-learning that can be used to address complex problems in optimization, classification, search, and other data-analytic contexts. EDA accounted for, or explained, 54% of the diagnostic variance in the sample data. Cardiovascular activity accounted for 34% of the difference variance in the guilty and innocent sample sampling data. The weighting coefficient for respiration was 12%. This weighting function is somewhat similar to other weighting functions in the polygraph literature. Although this study contributes little additional information to the published knowledge base, in addition to being computationally intensive and involving a small sample size, results of this study demonstrate the potential use for advanced computing techniques in polygraph research. Computing technology is more abundant and less expensive than in the past. Continued interest is indicated for both weighted EDA solutions, and the use of computational machine learning methods in polygraph research.

Polygraph testing, although often referred to conveniently as a lie detector, does not detect or measure lies, but instead relies on data that is primarily autonomic. These include respiration movement, electrodermal activity, cardiovascular activity and sometimes vasomotor activity. Analysis of polygraph data involves a series of functions similar to other data analytic contexts, including feature extraction, numerical transformation and data reduction, the use of some form of likelihood function, and structured decision rules to parse a categorical test result from the numerical and probabilistic data. An important challenge of any multivariate analysis is the calculation, or optimization, of a statistical function that specifies an optimal combination of the different sources of data that will achieve a desired objective.

Optimization refers to the calculation or computation of a best attainable solution. Optimization is a data-analytic approach to solution finding, as opposed to solution-finding through conjecture or anecdotal example, subjective opinion, or even expert opinion (equivalent to subjective opinion and conjecture). One way to determine the optimal structural combination of sensor data will be to test every possible combination. However, attempting to test every possible solution will be an expensive and time-consuming expedition. The number of possible weighting coefficients or structural combinations of respiration, EDA, and cardiovascular is potentially infinite. To gain insight into the possibilities, if weighting coefficients are regarded as normalized decimal proportions (summing to 1) there are 166,650 possible combination using only two decimals of precision. With the addition of a fourth recording sensor (i.e., vasomotor), the number of possible structural functions will be 4,082,925. Three decimals of precision would increase the possible combinations exponentially, though with potentially little benefit.

Another method to optimize the structural combination of respiration, EDA, and cardiovascular activity (or any combination of response features) would be to use traditional statistical methods such as linear discriminate analysis, linear regression or logistic regres-

sion. A more modern approach to optimization and classification problems (also search problems, and prediction problems) is to use statistical learning theory (Hastie, Tibshirani & Friedman, 2009; James, Witten, Hastie & Tibshirani, 2013;), also referred to as machine-learning (ML) and artificial-intelligence (AI).

An important difference between AI and the traditional statistical approach is that the traditional approach involves a researcher or scientist who develops a hypothesis (possible solution) about a possible answer to a research question. The researcher then designs an experiment to falsify the hypothesis or to compare the hypothesis and null hypothesis to determine which is more consistent with the observed data. The AI approach allows a computing machine to both suggest and test numerous possible hypotheses. Thus, the machine is said to "learn" a solution from its experience with the data.

This project involved the use of a genetic algorithm (Goldberg, 1989; Mitchel, 1996) to compute structural combinations of respiration, EDA and cardiovascular activity. The optimization question is this: what is the best structural weighting for data from each of the polygraph sensors? In this context, *best* is defined as achieves the greatest number of correct decisions when classifying the sample cases as deceptive or truthful.

Data

Data consisted of a small sample of n=18 confirmed deceptive and n=18 confirmed truthful polygraph cases. The sample cases were conducted with a diagnostic polygraph format with two relevant questions. Cases were conducted by a large metropolitan police agency, consisted of respiration, EDA and cardiovascular activity data, and were confirmed through a combination of confession and extra-polygraphic evidence. Examinees were criminal suspects who authorized the examination, including the use of the data in anonymous form for research, program evaluation instruction and quality control. All exams consisted of three iterations (three charts) of the sequence of the test questions. All examinations consisted of sensors for thoracic and abdominal respiration movement, EDA, cardiovascular activity and an activity sensor.

The two relevant question diagnostic format is used for event-specific diagnostic polygraphs. It includes two relevant questions and three comparison questions, along with other procedural questions. When using the two relevant question diagnostic polygraph format, each relevant question is evaluated with the preceding or subsequent comparison question depending on which comparison question has produced the greater change in physiological activity. All exams were conducted and recorded using the Lafayette LX4000 polygraph instrument.

Data were exported from the proprietary binary file format to the NCCA ASCII text format using a data sampling rate of 30 samples per second. Data were then imported to the R Language and Environment for Statistical Computing (R Core Team, 2019) for analysis. All feature extraction, numerical transformation, data reduction, likelihood calculations and decision rules were executed automatically in the R computing environment. The respiratory feature of interest was the reduction of respiration activity in response to the test stimuli, associated with attempts to conceal one's deception. The EDA feature of interest was the change in y-axis value from an onset of a positive slope segment to the peak of reaction, associated with increased activity in the sympathetic division of the autonomic nervous system. For cardiovascular activity data the feature of interest was the change in y-axis value, also associated with relative blood pressure and activity in the autonomic nervous system.

Feature extraction was performed for each sensor for each relevant question (RQ) and each comparison question (CQ). Respiration data was measured as the mean of respiration line excursion (RLE; the absolute difference of each subsequent respiration sample) for a one-second moving average from stimulus onset to 15 seconds post stimulus onset excluding the data from one second before to one second after the recorded verbal answer. This measurement is thought to be more robust against distortions at the point of verbal answer and is not influenced by the length of the 15 second evaluation window – effects with different measurement periods will have a

similar metric. EDA reactions were measured as the onset of a positive slope segment during a response onset window (ROW) from .5 seconds after stimulus onset to 5 seconds after the verbal answer to the greatest y-axis (vertical) distance to subsequent peak of reaction (onset of negative slope) within evaluation window (EW) from stimulus onset to 15 seconds after stimulus onset. If there was no response onset during the ROW a response onset was inferred statistically during positive slope segments using a z-test of the variance of one second mean difference of each subsequent EDA sample. A response onset was imputed if the difference in variance for a two, one-second windows exceeded the alpha = .001 boundary. This can be visualized as a substantial increase in positive slope angle within a positive slope segment during the ROW. Cardiovascular activity was extracted by first calculating the mean of all cardio sensor samples.

This can be thought of, and plotted, as the mid-line between the systolic and diastolic peaks. Cardiovascular activity changes were then extracted, using the cardio mid-line, using a procedure similar to the one for the EDA data.

All measurement values were dimensionless. That is, they were not indexed to any physical quantity, SI unit , or derived measurement value. Dimensionless values were then transformed to objective ordinal rank values using a three-point coding scheme [-1, 0, +1] familiar to field polygraph examiners. For each of the recording sensors, extracted values for each presentation of each RQ was compared to the preceding or subsequent CQ depending on which CQ produced the greater change in physiological activity. Scores were coded as +1 if the change in physiological activity was greater at the CQ and were coded as -1 if the change in physiological activity was greater at the RQ. Tied values (tied ranks) were coded as 0. For EDA and cardiovascular activity, a greater extracted value was indicative of a greater change in physiology. However, because the respiratory feature of interest involved the reduction of respiration activity, sign values were inverted so that smaller extracted values were interpreted as a greater change in physiological activity.

Non-parametric rank values were then reduced to subtotal scores for each RQ through summation. Subtotal scores were then summed to achieve a grand total score for each exam. The analytic theory of the polygraph test postulates that greater changes in physiological activity are loaded at different types of test stimuli as a function of deception and truth-telling in response to relevant target stimuli (Nelson, 2015, 2016). Under this theory, grand total scores of this type can be expected to be greater than zero for innocent examinees and less than zero for guilty examinees. The genetic algorithm was used to determine the weighting coefficients that can be assigned to scores from each of the recording sensors to maximize the number of correct classifications.

Analysis

A genetic algorithm can be thought of as a Monte Carlo method, involving the use of random numbers to create numerous possible solutions to a question or analytic problem. [See Eckhardt (1987), Metropolis, (1987), and Metropolis and Ulam (1949) for more information on Monte Carlo methods]. A genetic algorithm consists of simple rules such as the following:

1. Creation of numerous (say, m=1000) random possible solutions for the structural weighting of respiration, EDA and cardiovascular activity data,

2. Testing the effects of each possible solution with all of the sample cases,

3. Survival of the best solutions (natural selection) – discard the 50% that performs weakest and keep the 50% that achieves the best classification,

1 International System of Units (French: Système international d'unités, abreviated as SI). SI base units include the following: the meter as a measurement of length or distance, the kilogram as a unit of mass, the second as a unit of time, the ampere as a unit of electric current, the kelvin as a unit of temperature, the candela as a unit for luminosity, and the mole as a unit for the quantity of a substance. All other measurement units are derived from these SI base units. Measurement of any quantity requires both a physical quantity to measure and a defined unit of measurement

4. Split each of the surviving solutions into two parts and randomly connect them (recombination) to make a new iteration of m possible solutions for the structural weighting of the sensor data – now informed by the previous experience,

5. Introduce random variation (mutation) to a small portion of the new solutions – to potentially find better solutions that were not included in the previous solutions,

6. Repeat steps 2-5 a large number of times,

7. Stop at some point – either after a specified number of iterations (say, 30,000), or in response to the achievement of a stated objective (e.g. a desired level of accuracy), or when the structural model stops improving, and finally,

8. Choose the structural solution that achieves the greatest effect size

Results

The genetic algorithm used objective integer-level rank order input data and produced the weighting function shown in Table 1. EDA accounted for or explained over half of the diagnostic variance in the sample data. Cardiovascular activity accounted for approximately one-third of the difference between the guilty and innocent sample sampling data. Respiration data explained slightly over 10% of the diagnostic variance. This weighting function is somewhat similar to other weighting functions in the polygraph literature, including the discriminate function reported by Nelson, Krapohl and Handler (2008) in the development of the Objective Scoring System Version-3, also shown in Table 1.

Discussion

This project involved the use of a balanced sample of n=36 field polygraph exams and a simple genetic algorithm to compute a weighting function for polygraph signals that would optimize the classification of deception and truth-telling. A genetic algorithm, and other ML techniques, can achieve a very close approximation of an optimal solution with only a few thousand (sometimes many thousand) iterations. Response features in this study were coded with an objective rank method using positive and negative values [-1, 0, +1] by comparing responses to relevant and comparison stimuli. Input data were intentionally naive as to the relative importance of the data from different recording sensors, and the algorithm output is a weighting function that will optimize the diagnostic variance of the extracted data.

Table 1. Weighting function.

Sensor	Normalized weighting function	
	Genetic Algorithm	OSS-3
Respiration activity	.12	.19
EDA	.54	.53
Cardiovascular activity	.34	.28

EDA data accounted for over 50% of the variance while cardiovascular data accounted for approximately 1/3 of the diagnostic variance. Respiration data accounted for the smallest portion of diagnostic variance. This weighting coefficients are similar to other published information. Some manual scoring protocols approximate this weighting function by doubling EDA scores.

The procedures in this study differ from those commonly used scoring in field polygraph programs, in which manual/visual feature extraction continues to be a dominant method for the interpretation of polygraph test

data. It also differs from most studies on automated algorithm development in its use of ordinal integer-level numerical coding. Results from this study add additional confirmation to existing knowledge on the relative importance of polygraph signals, and may be helpful to better understand polygraph scoring methods such methods such as the OSS (Krapohl, 2002; Krapohl & McManus, 1999) and ESS (Nelson, Krapohl & Handler, 2008; Nelson et al., 2011; Nelson 2017).

Limitations of this project include the small sample size, and the limited information available about the case confirmation. Despite the sample size, results from this study appear to be consistent with other information on the structural weighting of polygraph signals. Another, limitation of this project, related to the small sample size, is the absence of a hold-out sample. No attempt was made, during this project, to test the effectiveness of the weighting function with other data. Also, no attempt was made to test the effectiveness of the weighting function with the study input data, as doing so would incur a risk of over-fitting a conclusion with the small input sample, and thereby overestimating its effectiveness. Another potential limitation, related to the use of Monte Carlo methods with small sample sizes, is that replication of these results may be subject to both sampling variation and Monte Carlo variation. This limitation is mitigated by the results of other studies on signal weighting in manual scoring methods – such as those already cited, the one by Nelson and Handler (2018) – that demonstrate the effects of weighting the EDA data more than the other sensor data. A final limitation of this study is that it is computationally intensive. However, computing power is much more abundant and much less expensive than in the past. Thoughtful use of computing and analytic technologies can help to improve and advance the science and field practice of polygraphic credibility assessment testing.

In consideration of the volume of existing information, results of this study are not surprising, and the results of this study contribute little new knowledge to the science and field practice of polygraph testing. Optimization of respiration, EDA and cardiovascular activity has previously been demonstrated using a variety of methods, including logistic regression and discriminate analysis and other methods. Monte Carlo methods have been described in previous polygraph studies. These results are interesting because they serve to add further confirmation of extant knowledge regarding polygraph signals, and it introduces and demonstrates the potential use of ML/AI techniques in polygraph studies. Continued interest is indicated for both weighted EDA solutions, and the use of computational machine learning methods in polygraph research.

References

Eckhardt, R. (1987). *Stan Ulam, John von Neumann, and the Monte Carlo method.* Los Alamos Science (15), 131–137.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley Professional.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with applications in R.* Springer.

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

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

Kroese, D. P., Brereton, T., Taimre, T., & Botev, Z. I. (2014). Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics* (6), 386–92.

Metropolis, N. (1987). The beginning of the Monte Carlo method. Los Alamos Science (*1987 Special Issue dedicated to Stanislaw Ulam*), 125–130.

Metropolis, N., Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association,* 44(247), 335–341.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms.* Cambridge, MA: MIT Press.

Nelson, R. (2015). Scientific basis for polygraph testing. *Polygraph 41(1)*, 21-61.

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. & Handler, M. (2018). Reducing inconclusive results: a descriptive analysis of decision rules, weighted electrodermal scores and multinomial cut-scores. *Polygraph & Forensic Credibility Assessment, 47 (2)* 108-121.

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.

R Core Team (2019). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

PCSOT Instant Offense Polygraph Exams:

A Comparison of the Two-Question and Four-Question Test Formats

Erika Thiel[1] and

Raymond Nelson[2]

Abstract

The Center for the Treatment of Problem Sexual Behavior in Connecticut administers post-conviction sex offender testing to offenders who are on state and federal probation or parole. Over the years the examiners have switched the methods they use to administer the instant offense exam. All exams in this comparison were scored with the Empirical Scoring System or the Empirical Scoring System, Multinomial. The methods that were used were either a 2- question diagnostic test, or a 4-question diagnostic test. Each method used grand total scoring per series. Outcome comparisons for the 2-question diagnostic test and the 4-question diagnostic test showed a reduction in the number of inconclusive results by 55%, a difference that was statistically significant ($p = .006$). The reduction of inconclusive results observed in this analysis provides support for the hypothesis that the 4-question test is a more powerful alternative than the 2-question test. It is suggested that polygraph examiners consider administering instant offense testing use a 4-question diagnostic test with a grand total scoring method to reduce their inconclusive rate.

The Center for the Treatment of Problem Sexual Behavior (CTPSB) is a part of The Connection Incorporated (TCI) in Connecticut. The program focuses on the treatment of sexual offenders who are on state or federal probation, or state parole. The program takes a collaborative approach in the treatment of sexual offenders (Center for Sex Offender Management, 2000). This is an inclusive process between supervising officers, clinical staff, the polygraph examiners, and other involved professionals. The collaborative approach allows all parties to be involved and coordinate with each other in the treatment, supervision and behavioral monitoring of persons convicted of sexual offenses, with the goal of promoting both healthy living and community safety.

[1] Erika Thiel became the Polygraph Manager of CTPSB in July 2015 and is employed in this position to date. She is also an elected member of the APA Board of Directors.

[2] Raymond Nelson research specialist with Lafayette Instrument Company (LIC) and an elected member of the APA Board of Directors.

Acknowledgments: This work was stimulated by the earlier works of Dr. David Raskin and his University of Utah colleagues. The authors are indebted to Dr. Raskin and Mr. Don Krapohl for reading this and earlier drafts of this paper.

The views expressed in this article are solely those of the authors and do not necessarily represent those of The Connection Inc, CTPSB, the APA, or LIC. There are no proprietary interests in connection with this project.

The instant offense polygraph is a type of polygraph used in post-conviction sex offender testing (PCSOT). The instant offense polygraph is given to clients who are denying all, or part, of the offense for which they were convicted or pleaded guilty. Instant offense polygraphs are often given to clients when they first begin treatment. In treatment, the results of an instant offense polygraph can help the clinical staff assist the client through their denial barriers. For supervising officers, the result of the instant offense polygraph can help determine a referral and placement, within treatment and within the supervising agency, that is appropriate to the individual's needs. In 2015, the polygraph team analyzed the usefulness of the Empirical Scoring System (ESS) and began using ESS (Nelson & Handler, 2010; Nelson & Handler, 2012; Nelson, Handler, Shaw, Gougler, Blalock, Russell, Cushman & Oelrich, 2011). All exams have been analyzed with the ESS since that time. Later training and program development activities led to additional changes with the polygraph test format used with instant offense polygraph.

Method and Results

In July 2015 the polygraph team was using a two-question diagnostic test format. Clients were administered multiple series, during one exam, if there were multiple aspects that the client was denying about their offense. For example, series one may have focused on direct physical sexual contact, series two may have focused on verbal force or resistance, and series three may have focused on physical force or resistance. There were never more than three series administered in an attempt to avoid testing fatigue and unusable data. Each series used grand total scoring to determine the outcome for that series. All information and results were reported in the polygraph report that was shared with clinical staff and supervising officers.

Prior publications described a four-question diagnostic exam with grand totaling (Nelson, 2018; Raskin, Honts, Nelson & Handler, 2015; Raskin & Honts, 2002; Raskin & Kircher, 2014). Published information suggested this may be a more powerful test format than the two-question diagnostic test.

Starting September 1, 2016, the polygraph team began to administer the four-question diagnostic exam for the instant offense polygraph. The questions would focus on the totality of the offense and details of the allegation that the client was denying. This means that in one series the questions may focus on direct physical contact, verbal force or resistance, physical force or resistance and any other aspect. When using the four-question diagnostic format these questions are treated as non-independent. That is, it is expected that past behavior that could affect responses to one question may also influence responses to other questions. If there were not four different testable aspects of denial, then questions were formulated using a "who," "what," "where" and "when" approach. In switching to this method, there has yet to be an issue with coming up with four questions based on the client's denial and the victim statement.

The Empirical Scoring System, Multinomial (ESS - M; Lafayette, 2018; Nelson, 2017a) was updated in the Lafayette Instrument Software (version 11.8). The polygraph team updated the software and switched to ESS - M on September 17, 2018. There were no changes made to the default preferences in the system. The team changed their cut scores in accordance with the scoring method. ESS cut scores for grand total scores using the two-question test were alpha = .05 for deception (-4) and alpha = .1 for truth-telling (+2). For ESS-M the numerical cut scores for grand total scores are alpha = .05 for deception (-3) and alpha = .05 for truth-telling (+3). Confidence intervals reported by Nelson (2017b) showed no significant differences in decision accuracy or inconclusive outcomes for the ESS and ESS-M. There were no other changes made to the administration or analysis of the four-question diagnostic exam for the instant offense.

Table 1 below show the outcomes per test type and scoring system. The two-question testing dates were from July 2015 through August 2016. The four-question testing data collection dates are September 2016 through August 2019.

Table 1- Categorical results for the two-question and four-question Instant Offense Testing

Two Question Diagnostic Exam with Grand Total Scoring: ESS		
No Deception Indicated (NDI)	69	26%
Deception Indicated (DI)	139	52%
Inconclusive (INC)	28	11%
Unusable Date	29	11%
Total Tests	265	
Four Question Diagnostic Exam with Grand Total Scoring: ESS		
NDI	81	21%
DI	260	68%
INC	20	5%
Unusable Data	23	6%
Total Tests	384	

Without knowledge of the case status, statistical comparison of DI and NDI results is potentially misleading, as these outcomes can be assumed to be non-random (if the theory and practice of polygraph are valid). Observed outcomes will be informed by prior base-rates which may be influenced by systemic factors within the judicial system, probation and parole systems, and treatment system. In other words, observed differences in DI and NDI could possibly result from differences in referrals (i.e., differences in examinees), or other programmatic factors (e.g. judicial accuracy, or clinical resolution of denial), during the time periods of data collection. These differences might also affect the observed rate of inconclusive results.

Previous studies have not suggested differences in ESS and ESS-M accuracy. Testing the level of significance of observed differences requires an ability to characterize the difference as random under the null hypothesis. Statistical comparison of unusable data is also not a completely reasonable topic for a statistical hypothesis test because the rationale involves insufficiently differentiated factors as to why the change in test format would interact with this experience. Inconclusive results – not due to unusable data – may be thought of as related to differences in the polygraph test format with the null-hypothesis that there is no difference. Exact reasons for inconclusive results are unknown and therefore uncontrolled. We therefore assumed, under the null hypothesis, observed differences in inconclusive rates can be characterized as random. Whereas uncontrolled variation represents one degree of freedom, attempts as analysis or interpretation of NDI and DI results involves multiple degrees of freedom (e.g., TP, FN, FP and FN rates in addition to the unknown base rate or criterion state of each case). For these reasons, statistical tests were performed only on the inconclusive outcomes.

A bootstrap hypothesis test of 100,000 iterations showed that the observed difference in inconclusive rates was statistically significant ($p = .006$). The four-question diagnostic test had significantly fewer inconclusive results than the two-question diagnostic test.

Discussion

Changing the polygraph testing format from a two-question diagnostic exam with multiple series to a four-question diagnostic exam with one series reduced the number of inconclusive results by 55% - a significant reduction. A reduction of 46% was also observed in the number of tests that produced unusable data - though it is not completely clear as to why this occurred. Another analysis of outcome production should be done in another year when more information becomes available.

From the results listed, a four-question diagnostic reduces the proportion of inconclu-

sive results in comparison to a two-question diagnostic test. This is possibly due to better target selection when asking four questions as opposed to two. It is also possibly due to the fact that the four-question diagnostic format provides more data with which to classify the results as deceptive or truthful. The results also show a decrease in the amount of times unusable data was collected. This is possibly due to the fact that there are fewer test charts and therefore less test fatigue. It may also be speculated that a four-question diagnostic may lessen the chance of a false negative results due to the test fatigue and the exposure to practicing the test by running multiple series with a two-question diagnostic test. Because the rationale is speculative and as to why an effect for unusable data would interact with the test format, no statistical test was conducted on this observed difference.

A limitation can be noted at this point – in that to the extent that referral (examinee) differences and program differences, during the two time periods of data collection, may have influenced DI and NDI outcomes, these differences may have also played a role in the observation of inconclusive results. Another, limitation of this study is that it is unable to attribute the observed differences in inconclusive results to significant improvements in test sensitivity, specificity or both. Regardless of this limitation, these results may be of interest to other professionals. To the extent that inconclusive test results are the result of uncontrolled (i.e., random) variation, The reduction of inconclusive results observed in this analysis provides support for the hypothesis that the 4-question test is a more powerful alternative than the 2-question test. These results should be compared with the results of other studies.

The polygraph examiners of CTPSB have assessed their data and will continue the use of the four-question diagnostic test for instant offense polygraphs with the ESS-M scoring method to collect further data for analysis. Based on the experience of this team, it is suggested that a four-question diagnostic test may help reduce the rate of inconclusive results for other examiners administering instant offense exams in PCSOT programs. Future studies should attempt to compare the sensitivity and specificity rates of these polygraph formats in addition to their inconclusive rates.

References

Center for Sex Offender Management. (2000). The Collaborative Approach to Sex Offender Management. Available online at [https://www.csom.org/pubs/collaboration.html].

Lafayette Instrument Company (2018). Empirical Scoring System, Multinomial.

https://lafayettepolygraph.com/ess-11_8.pdf.

Nelson, R. (2018). Credibility assessment using Bayesian credible intervals: a replication study of criterion accuracy using the ESS-M and event-specific polygraphs with four relevant questions. Polygraph & Forensic Credibility Assessment 47(1), 85-90.

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

Nelson, R. (2017b). Updated Numerical Distributions for the Empirical Scoring System: An Accuracy Demonstration with Archival Datasets with and without the Vasomotor Sensor. Polygraph & Forensic Credibility Assessment Testing 46(2), 116-131.

Nelson, R. & Handler, M. (2010). Empirical Scoring System. Lafayette Instrument Company.

Nelson, R. & Handler, M. (2012). Using Normative Reference Data with Diagnostic Exams and the Empirical Scoring System. APA Magazine, 45(3), 61-69.

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.

Raskin D.C. & Honts, C.R. (2002). The comparison question test. In M. Kleiner (Ed.), Handbook of polygraph testing. London: Academic (1 – 49).

Raskin, C. R. Honts, & J. C. Kircher (Eds.), Credibility Assessment: Scientific Research and Applications. San Diego, CA: Elseveir/Academic Press.

Raskin D. C., Honts, C. R, Nelson, R., & Handler, M. (2015). Monte Carlo Estimates of the Validity of Four Relevant Question Polygraph Examinations. Polygraph, 44 (1), 1-27.

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. 63-129). San Diego, CA, US: Elsevier Academic Press.

A Brief Discussion of the Lower Latency Limit of the Electrodermal Response in Polygraph Testing

Donald J. Krapohl[1]

Donnie W. Dutton[2]

Karen A. Nix[3]

Abstract

Polygraphy and the larger field of psychophysiology share an interest in extracting meaningful information from bodily signals. The electrodermal response is one of those signals, and one of the most researched. Best estimates suggest that in polygraphy, the electrodermal response accounts for about half of the diagnostic information available in the charts, and understanding its characteristics is crucial to accurate analysis by polygraph professionals (Kircher & Raskin, 1988). In a recent instruction article, Krapohl and Nix (2019) challenged a common assumption among many polygraph examiners that the minimum latency of an electrodermal response in polygraph testing was 0.5 seconds. Here we summarize specifically the current state regarding electrodermal response latency. We continue by discussing methods for standardizing question presentations to permit polygraph examiners the ability to enjoy a higher reliance on latency information.

Introduction

Electrodermal activity is used in both psychophysiological research and polygraph testing. Its utility stems from the generally established and accepted conclusion that EDA provides a reliable proxy for arousal (Boucsein, 2012), and that electrodermal response amplitude covaries with the arousal value of a stimulus, though not linearly.

There are three characteristics of a stimulus that will induce a phasic electrodermal arousal: novelty, intensity, and salience (Dawson, Schell & Filion, 2007). Salience of a stimulus can have any number of sources, including those that are associated with the act of deceiving. Polygraph testing entails the manipulation of the salience of test items, and because arousal corresponds with salience, patterns of arousal permit inferences regard-

[1]Former APA President and Editor, currently with the Capital Center for Credibility Assessment.

[2]Former APA President and current Director. Mr. Dutton is Vice President of the Capital Center for Credibility Assessment.

[3]Certified polygraph examiner through the South Carolina Law Enforcement Division, currently with the City of Charleston (SC) Police Department.

This article is another in the Best Practices series.

The authors are grateful for thoughtful comments and suggestions from Mark Pszenny, Dr. Ian Dersley, Pamela Shaw and APA Editor Mark Handler. The opinion expressed in this article are those of the authors and do not necessarily represent the views of past or current employers, or the APA. The authors have no financial interests in any product or service related to this article. Questions and comments can be sent to the first author at APAkrapohl@gmail.com.

ing the salience value of the test questions. Test questions that challenge an examinee's goal of passing the test may garner salience in this way (Khan, Nelson & Handler, 2009).

An additional value of using electrodermal activity (EDA) is that the underlying mechanisms are well understood by scientists. A thorough summary of electrodermal activity as it applies to polygraphy was published in a previous American Polygraph Association (APA) journal by Handler, Nelson, Krapohl, & Honts (2010).

All physiological responses have some period of delay between the stimulus and the beginning of the reaction. This is true of EDA, as well. Estimates of electrodermal response (EDR) latency can be traced to some of the earliest published articles on this phenomenon at about 1 – 3 seconds (Tarchanoff, 1890, as cited by Peterson & Jung, 1907).

Our present interest regards the lower limit of EDR latency in polygraph testing: How early can an EDR begin after the onset of the question and be reliably associated with the content of that question? EDR minimum latency matters in polygraphy, just as it does in psychophysiological research. EDRs must be excluded from scoring if they clearly could not come from the test stimuli, calling for rules about what constitutes a timely response.

There is a prevailing assumption in polygraphy that EDRs beginning 0.5 seconds or later after question onset can be associated with the content of the test question, and therefore are scorable. The documentary trail leads to Raskin and his collaborators at the University of Utah as the source. The earliest articles we found to suggest 0.5-second latency for polygraph were by Podlesny and Raskin (1978), and Raskin (1979). The first mention of the 0.5-second minimum EDR latency to appear in the polygraph practitioner literature was 20 years ago, by Bell, Raskin, Honts & Kircher (1999). Raskin and his collaborators have long been highly influential in the field of polygraph for their substantial contributions to evidence-based practices, and as such the 0.5-second latency gained widespread acceptance. It did not only affect manual polygraph scoring: 0.5-second EDR latency is programmed into the OSS-3 algorithm (Nelson, Krapohl & Handler, 2008) as well as the CPS algorithm (Kircher & Raskin, 2000).

The most widely read polygraph textbooks offer inconsistent minimum periods for EDR onset latency. Abrams (1992) suggests any response after the first few words of the test question can be used. Reid and Inbau (1977) and Matte (1996) did not offer a fixed value. Krapohl and Shaw (2015) repeat the 0.5-second minimum latency. The 2017 version of the National Center for Credibility Assessment (NCCA) pamphlet on scoring states "3.6.2. The response onset window for the ED (electrodermal) and CV (cardiovascular) channel is from stimulus onset to five seconds beyond the examinee's answer" (parentheticals added). Even among those who did offer recommendations for minimum EDR latency, none cited the evidence on which their recommendations were based.

We undertook a literature review to the question we posed about EDR onset latency and found it has been well researched over the past 50 years. There did appear to be substantial evidence on which to base an EDR onset latency minimum. The research simply had not been translated into practical advice for polygraph practitioners. The limits of the nervous system and the sweat glands, and the type of stimuli used in polygraph testing make a strong case that the 0.5-second EDR latency standard is untenable. The following provides the basis for this assertion.

Brief Physiological Foundation

In the central nervous system, electrodermal activity (EDA) is largely mediated by the hypothalamus, a brain structure principally responsible for the sweating response to both thermoregulatory requirements and emotional arousal among many functions, though there are also other less impactful contributors (See Boucsein, 2012). Signals from the hypothalamus travel via sympathetic pathways to preganglionic sudomotor neurons. From there the signals are transferred through postganglionic neurons to the sweat glands. The sweat glands release fluids through ducts towards the skin's surface. Sweat is a good conductor of electricity, and momentary fluctuations in sweat availability on the skin's surface change

the conductance of electricity, and through circuits and filtering produce the physiological channel called EDA.

The latency of EDRs depends in large part on how long it takes to communicate the signal from the brain to the eccrine glands, and for the eccrine glands to move sweat to the surface of the skin. Compared to other neurons, the nerve fibers that serve to communicate the electrical signal to the sweat glands are relatively slow. The best estimate is that it takes about 1.1 seconds for a nerve impulse to travel from central activation to the eccrine glands at the fingers (Lim, Seto-Poon, Clouston & Morris, 2003). At the sweat glands there is also a relatively long latency in the neuroeffector transfer to the sweat glands of 0.348 seconds (Kunimoto, Kirnö, Elam & Wallin, 1991). This produces a total average time of about 1.5 seconds from the activation of the EDR signal in the brain to the initiation of movement of sweat out of the eccrine glands. We were unable to locate a time estimate for the flow of sweat from the eccrine sweat glands to the measured decrease in resistance at the recording site. Ignoring this period and that for processing the information in the brain to initiate the response, the intervening physiology is responsible for an average of about 1.5 seconds of EDR latency. Edelberg (1972) put a bottom limit of EDR latency at 1.2 seconds for even the fastest responders. Similar minima have been reported using different methodologies (Barry, 1990; Lockhart, 1972; Surwillo, 1967; Venables & Martin, 1980).

More recently, using state-of-the-art computerized equipment with exquisite temporal resolution, Sjouwerman and Lonnsdorf (2019) found average EDR latency of 1.92 seconds to auditory startle prompts with a sample of 281 young adult participants. They did not report the minimum latency, but their graphed data suggest a lower limit near 1.3 seconds for a limited number of EDRs from the fastest of their research participants. Sjouwerman et al. (2019) concluded their findings were consistent with those of earlier researchers using analog instruments to investigate EDR latency. Startle prompt-induced EDRs have the shortest onset latencies a person can produce (Sjouwerman et al., 2019).

Variations in latency minima may result from differences in transit distances from brain to fingers, the subject's age, health, skin temperature (Boucsein, 2012), and whether the stimuli are visual, auditory or tactile (Sjouwerman et al., 2019. Because polygraphy almost always uses auditory presentations of test questions we have restricted this review to addressing auditory stimuli.

Polygraph Stimuli

Startle prompts (e.g., bursts of 105 dB white noise) require no cognitive interpretation to elicit EDRs. Not so with polygraph questions. In polygraphy, the interest is not the reaction to the physical sound of the test question, but how the examinee reacts to the semantic content of the question. The EDR delay, then will be influenced by how quickly the question is presented and how many words of the question have been given before the examinee has sufficient information to know what the question is. At what point the examinee comes to know the meaning of the question may vary by examinee and question.

Polygraph test questions are typically about 3 to 20 words in length. All questions are normally rehearsed with the examinee at least one or more times before testing to ensure the examinee understands their meaning. The review process may also prime the examinee to questions having more personal salience.

As stated previously, to respond to the meaning of the test questions during testing the examinee must hear enough of the question to know what the question is, or at least what kind of question it is. In some cases, an examinee can recognize the kind of question having heard only the initial word. Other cases may require more words for the examinee to know which question it is, such as when relevant question covering different issues share an initial phrase like "Apart from what you told me.." or "In the past two years have you…", or the more lengthy "Other than what you told me, in the past two years have you…". If the average person speaks at a rate of three words per second, examinees may come to ascertain the meaning of a test question from 0.3 to more than 4 seconds after the question begins. After processing the ques-

tion, the central nervous system may initiate a signal. Once initiated, the autonomic nervous system takes an average of 1.5 seconds to induce the sweat glands to release sweat to the skin surface, as the previous section discussed. With few disputable assumptions the average EDR would be expected to begin more than 1.8 seconds from question onset for the average examinee, with some individual differences. Other factors, such as cooler skin, can be expected to delay it further (Maulsby & Edelberg, 1960).

Based on the available evidence, the average minimum EDR onset latency for polygraph questions should be more than 1.8 seconds for most healthy and responsive examinees who recognize the test question at the presentation of the first word of the question. This estimate approximates the Sjouwerman and Lonnsdorf (2019) findings of an average of 1.92 seconds latency for auditory startle prompts. Of course, these are only averages and assume immediate recognition of the test question. Many individuals would be expected to have longer, and sometimes substantially longer EDR onset latencies. Even under optimal conditions, human physiology constrains EDR onsets for the fastest responders to not less than 1.2 seconds, however, EDRs taking place this quickly would be statistically rare.

Problems of Latency Measurement in Polygraph Testing

A casual inspection of field polygraph charts may lead to the conclusion that EDR latency can be shorter than one second, contrary to the established limits of the nervous system. We suggest there are two alternate explanations for such a finding, both due to human factors.

The first likely source of short latencies could be when examinees can see or hear their examiner preparing to ask a question. If examiners unconsciously clear their throats, take an audible deep breath, type on the keyboard, or shift in their chairs just before asking test questions, examinees can hear the examiner and initiate an anticipatory response without having heard any of the test question. That early reaction would be due to an extraneous noise, not because the examinee knows which question is about to come. Similarly, human visual fields are about 210 degrees, sometimes permitting examinees to see their examiners moving their hands to the computer keyboard as they prepare to timestamp the presentation of the question. Again, an early reaction from the examinee can be triggered by the examinee being alerted that something is about to happen. In these instances, examiners themselves may unwittingly create premature EDRs.

The second source is the differing ability of examiners to mark the polygraph recordings when a test question is asked. Examiners do not always precisely match the onset of their speaking with the key press that timestamps the beginning of the test question in the data. Consequently, latency information can be distorted. It can be difficult to determine whether a very short EDR onset period is real or merely a reflection of the examiner's error in marking the true question beginning.

Neither of these problems exist in modern day research settings because scientists take care to automate the stimulus presentations and timestamping. These two measures ensure standardized stimulus presentations and eliminate a source of recording error. It is not as common however, for polygraph examiners to undertake the same safeguards. Currently most examiners read the test question to the examinee themselves and mark the question presentation in the data manually. The polygraph method for verbal question presentation and manual event marking has not changed since the 1920s. This is not due to a lack of opportunity to exploit automation: Almost all digital polygraphs already offer options for automated question presentation with accompanying event marking on the charts. These features offer a vastly improved method for standardization and event timing. For polygraph examiners to protect against misattributing premature EDRs to test questions they could automate question presentations. For additional protection, the use of audio headphones to present the test questions could block out extraneous sounds that can come from testing at sites where noises are otherwise difficult to control. And finally, automated test question presentation might be associated with increased decision accuracy (Honts & Amato, 1999).

Summary

Psychophysiological research points to a mean delay between the onsets of a startle stimulus and an associated EDR of just under 2 seconds. The shortest of EDR latencies cannot be shorter than 1.2 seconds under any conditions. Because polygraph testing uses stimuli that require a degree of interpretation beyond simple alerting, typical latencies might be expected to be longer in the polygraph context. We suggest the average minimum EDR onset for polygraph examinees should be slightly longer than about 1.8 seconds from the beginning of the question presentation. Because it is exceedingly difficult to determine precisely when polygraph examinees recognize the meaning of test questions, the maximum EDR latency is a question with no clear answer. It is reasonable that examinees vary in the speed in which they can recognize questions, and that individual differences might prohibit a fixed maximum onset latency that is appropriate for all examinees. An alternative approach may be to consider intra-examinee consistency in latency rather than inter-examinee averages. Nonetheless, it can be stated with confidence that EDR latencies shorter than 1.2 seconds cannot be elicited by the test question, and onset delays approaching this extreme should be rare.

Meaningful information regarding EDR latency in polygraph testing is hampered by the imprecise means in which stimulus onsets are typically established – from key presses by humans who concurrently attend to other tasks. A practical solution for polygraph examiners is to avail themselves of existing polygraph software to present the test questions and timestamp the data accordingly, leaving the examiner's attentional resources for other demands, such as monitoring the examinee. Examinees may react in anticipation if they see or hear the examiner preparing to present the test questions and examiners should adjust their procedures to avoid this possibility. The use of audio headphones to automatically present the test questions would have an additional benefit in the protection against external sounds that may elicit EDRs.

Polygraph examiners can have more confidence in the interpretation of EDRs if they have undertaken steps to ensure latency information has been faithfully recorded. Suggestions in this paper are offered to help avoid errors regarding an EDR's timing, and hence, its association with the test question.

References

Abrams, S. (1992). *The Complete Polygraph Handbook.* Lexington Books: Lexington, MA.

Barry, R.J. (1990). Scoring criteria for response latency and habituation in electrodermal research: A study in the context of the orienting response. *Psychophysiology*, 27(1), 94 – 100.

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

Boucsein, W. (2012). *Electrodermal Activity, 2nd Ed.* Springer: New York.

Dawson, M.E., Schell, A.M., and Filion, D.L. (2007). The Electrodermal System. In J. Cacioppo, L.G. Tassinary and G.G. Berntson (Eds.) *The Handbook of Psychophysiology*, *3rd Ed.* Cambridge University Press: New York.

Edelberg, R. (1972). Electrical activity of the skin. Its measurement and uses in psychophysiology. In N.S. Greenfield and R.A. Sternbach (Eds.) *Handbook of Psychophysiology*. Holt, Rinehart & Wilson: New York.

Handler, M., Nelson, R., Krapohl, D., and Honts, C.R. (2010). An EDA primer for polygraph examiners. Polygraph, 39(2), 69 – 108.

Honts, C.R., & Amato, S.L. (1999). *The Automated Polygraph Examination:* Final report of U.S. Government Contract No. 110224-1998-MO. Boise State University.

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

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

Kircher, J.C., and Raskin, D.C. (2000). CPS comments on Dollins et al.'s computer algorithm comparison. *Polygraph*, 29(3), 250 – 252.

Krapohl, D.J. & Nix, K. (2019). Evidence and common sense: Suggestions for scoring electrodermal responses. *APA Magazine*, 51(2), 60-76.

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

Kunimoto, M., Kirnö, K., Elam, M., and Wallin, B.G. (1991). Neuroeffector characteristics of sweat glands in the human hand activated by regular neural stimuli. *Journal of Physiology*, 442, 391 – 411.

Lim, C. L., Seto-Poon, M., Clouston, P. D., & Morris, J. G. L. (2003). Sudomotor nerve conduction velocity and central processing time of the skin conductance response. *Clinical Neurophysiology*, 114, 2172 – 2180.

Lockhart, R.A. (1972). Interrelations between amplitude, latency, rise time, and the Edelberg recovery measure of the galvanic skin response. *Psychophysiology*, 9, 437 – 442.

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

Maulsby, R.L., & Edelberg, R. (1960). The interrelationship between the galvanic skin response, basal resistance and temperature. *Journal of Comparative &Physiological Psychology*, 53, 475 – 479.

National Center for Credibility Assessment (2017, Aug). *Test Data Analysis: Numerical Evaluation Scoring System Pamphlet.* Ft. Jackson, SC.

Nelson, R., Krapohl, D., and 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.

Peterson, F., and Jung, C.G. (1907). Psycho-physical investigations with the galvanometer and pneumograph in normal and insane individuals. *Brain*, 30(2), 153 – 218.

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

Raskin, D.C. (1979). Orienting and defensive reflexes in the detection of deception. In H.D. Kimmel, E.H. van Olst and J.F. Orlebeke (Eds.) *The Orienting Reflex in Humans.* Erlbaum: Hillsdale, NJ.

Reid, J.E., and Inbau, F.E. (1977). *Truth and Deception: The Polygraph ("Lie-Detector") Technique*, 2nd *Ed.* William & Wilkins: Baltimore.

Sjouwerman, R., and Lonsdorf, T.B. (2019). Latency of skin conductance responses across stimulus modalities. *Psychophysiology, 56*(4), https://doi.org/10.1111/psyp.13307.

Surwillo, W.W. (1967). The influence of some psychological factors on latency of the galvanic skin reflex. *Psychophysiology, 4*, 223-228.

Tarchanoff, J. (1890). Galvanic phenomena in the human skin in connection with irritation of the sensory organs and with various forms of psychic activity. In *Pflüger's Arch. für Physiologie, 47,* 46 – 55.

Venables, P.H., & Martin, I. (1980). Electrodermal techniques. In I. Martin & P.H. Venables (Eds.), *Techniques in Psychophysiology* (pp. 3-57). Wiley & Sons: New York

Literature Survey of Structural Weighting of Polygraph Signals:

Why Double the EDA?

Raymond Nelson

Abstract

Electrodermal activity (EDA) is a useful and important source of information in psychophysiological detection of deception (PDD) testing. A variety of methods have been used to evaluate the relative contribution of EDA data, and results have been consistent throughout several decades of research in different laboratories. Information indicates that EDA data contributes approximately half of the information used to make effective classifications of deception or truth-telling in the CQT paradigm. This paper is a survey of existing literature on the correlation and weighting coefficients for EDA and other recorded data relative to PDD.

Introduction

Why double the electrodermal activity (EDA) scores in a polygraph setting? EDA is a useful and important source of information in psychophysiological detection of deception (PDD) testing. Scientific experiments are often used to study the strength of association between an unknown phenomena of interest and some observable data that serve as a proxy for the unknown phenomena of interest. In addition to studying the strength of association, test developers are also concerned with determining the optimal weighted combination of the measures/data that will maximize test effects such as test sensitivity, and specificity, or false-negative and false-positive errors. In the literature review that follows, summarized in Table 1, one can trace the development of the technology and study designs leading to the currently used weighting coefficients for EDA and other sensor data in the PDD testing context. Table 1 contains the studies in temporal order, the fitting technique type, the coefficients for the non-normalized data and the coefficients for the normalized data which leads to the current weighting system.

Methods

Published literature was surveyed for information relating measures of respiration, electrodermal (EDA), cardiovascular activity and vasomotor activity to psychophysiological detection of deception (polygraph) test outcomes. These signals have emerged as useful in the PDD testing context because they have been shown to be correlated with differences in deception and truth-telling, while contributing unique or additional variance to accurate classifications when used in combination. Criterion coefficients of interest, which describe the relationship between the data and the criterion state of deception or truth-telling, are summarized in Table 1. A number of different types of coefficients have been reported in various studies, including correlation coefficients, discriminate and regression coefficients and other structural coefficients. The optimal weights for each of the physiological measures was determined by the data from each of the included studies. The relative contributions to the optimal weights, as seen over time, converge to a scoring system with roughly half the weight given to EDA, i.e. twice that given to the other recorded signals.

Also included in Table 1 are the normalized coefficients. Normalization of values ensures that all information can be compared on a similar scale for which the sensor values will sum to one (1). The actual structural models can be expected to differ somewhat from this simplification. However, normalization in this manner can help readers to easily appreciate the relative strength of contribution of the different recording sensors.

Literature review

Kubis (1962) described an early study, funded by the U.S. Air Force, on the feasibility of using computers in polygraph testing. Discriminate functions were calculated for each of three polygraph test charts, the means of the three discriminate function are shows in Table 1. EDA accounted for over 60% of the diagnostic variance in that study. Kubis concluded that there was sufficient validity to the comparison question test – premised on the loading of physiological responses to relevant and comparison stimuli – to warrant confidence in the polygraph as a possible aide to a decision about whether or not to interrogate. Kubis noted the high degree of subjectivity and variability in manual polygraph feature extraction at that time, especially with regard to respiration and cardiovascular activity. Additionally, he commented on the weaker available computing technology at that time and recommended against attempts to develop computerized polygraph systems that would be intended to make on-the-spot decisions. Kubis recommended the use of computers in studying and developing polygraph features and polygraph feature extraction. One difference between the instrumentation used in this study and that of a modern polygraph instrument is that cardiovascular activity was recorded with an electronically amplified fingertip (mechanical circumference) plethysmograph, which is different than both the arterial pressure cuff and the photoelectric plethysmograph in use today.

Kubis (1964), reported the results of a second study on the feasibility of using computers in polygraph testing, also funded by the U.S. Air Force. This study also used a mechanical/circumference fingertip plethysmograph in lieu of the arterial cardio cuff sensor. Results of manual/numerical scores were reported along with the results of an objective measurement approach to feature extraction. Data shown in Table 1 are the concordance rate or frequency deceptive and truthful scores that concurred with the criterion states of the sample cases. Normalizing the proportions resulted in an EDA coefficient of .41, for summed numerical scores and .39 for discriminate scores.

Kircher (1981) used a sample of community participants to study the use of computers in the evaluation of PDD test data. Statistical and structural coefficients were reported for a variety of potential physiological measurements in PDD testing and analysis, including the three-chart means of standardized discriminate function coefficients for EDA, cardio and respiration data. Coefficients were included for both manual scores and computer/measurement scores. Computer scores also included a vasomotor sensor. EDA accounted for 50% of the variance in manual scores and 41% of the variance in computer scores for the guilty and innocent participants. Kircher concluded that the computer may be capable of offering improvements in polygraph data analysis.

Kircher (1983) reported the results of another study on the use of computers in lie detection and reported point-biserial correlation coefficients for both manual and computerized feature extraction with respiration, EDA, cardiovascular and vasomotor sensors. In that study, EDA accounted for 28% of the variance in manual scores and nearly 47% of the variance in computer scores for guilty and innocent participants. Kircher also reported the results of a discriminate function, for which the normalized coefficient for EDA was .46. Kircher reported accuracy ranging from 87.5% to 93.7%, with the computer algorithm slightly more effective at identifying programmed guilty subjects and manual scores slightly more effective at programmed innocent subjects, suggesting that physiological response patterns may be qualitatively different for deception and truth-telling. All point biserial coefficients in Table 1 are shown as r_{pb2} to facilitate more intuitive comparison of

results from different studies.

Kircher and Raskin (1988) reported a comparison of human and computerized evaluations of polygraph data, including respiration, EDA, cardio and vasomotor sensors. Point-biserial correlations are shown as r_{pb2} in Table 1. For manual scores the EDA accounted for 26% of the variance. However, EDA scores may be strongly covariant with vasomotor activity, also an indicator of autonomic activity, and the vasomotor scores also accounted for 26% of the diagnostic variance. For computer scores, for which the vasomotor sensor did not contribute additional information and was not included in the structural model, the EDA accounted for 46% of the observed variance in deceptive and truthful outcomes for programmed guilty and innocent subjects.

Raskin, Kircher, Honts and Horowitz (1988) completed a field study that was funded by the National Institute of Justice, and reported point-biserial correlation coefficients for respiration, EDA, and cardiovascular activity. Point-biserial correlations are shown as rpb2 in Table 1. EDA data alone accounted for 53% of the observed variance in deceptive and truthful outcomes for the guilty and innocent examinees in the field sample.

Raskin and Kircher (1990) completed a study on the use of computer algorithms in polygraph data analysis and countermeasure detection. They reported the coefficients for a discriminate function that included respiration, EDA, and cardiovascular activity. EDA accounted for 55% of the information in the normalized discriminate function that optimized the separation of guilty and innocent study participants. Discriminate coefficients and the normalization are shown in Table 1.

Capps and Ansley (1992) completed a survey of scores assigned by field examiners for each recording sensor and each scoring feature using a seven-point numerical scale. As shown in Table 1, over 40% of the points were assigned to the EDA sensor data.

Harris and Olson (1994), in a patent filing for the Polyscore algorithm, described the coefficients of a logistic regression that included EDA, cardiovascular activity, and respiration (also included pulse line length and pulse rate though these are not shown in Table 1). The EDA weighting coefficient was 49% after normalizing only the information that is most similar to traditional polygraph feature extraction (respiration, EDA and cardiovascular activity).

Ansley and Krapohl (2000) completed a survey of examiner scores using the seven-position scoring method with confirmed field cases. Table 1 shows that EDA scores accounted for 55% of all assigned scores in the sample data.

Honts, Amato and Gordon (2000) completed a study on the outside issue question. The study report included correlation coefficients for each of four scorers for respiration, EDA and cardiovascular activity, in addition to the vasomotor sensor. Table 1 shows the correlations in the form of a coefficient of determination (r^2). The coefficient of determination is an estimate of the observed variation in deceptive and truthful outcomes for guilty and innocent examinees that is explained by each sensor alone (without the addition of the other sensors). The mean of the r^2 coefficients for the four scorers is shown in Table 1. The coefficient of determination for EDA was .39. After normalizing the coefficients to compare their relative strength, EDA produced a coefficient of .41 when the vasomotor data was included and .57 when vasomotor data was excluded.

Kircher, Krisjianson, Gardner and Webb (2005) studied the validity of various scoring criteria that were previously taught at the U. S. government polygraph school and other accredited polygraph training programs in the past. This study concluded that some of the scoring features in the past were unnecessary – leading to a reduction of scoring features by the Department of Defense (2006) to only those features that are supported by scientific evidence – and suggested that primary scoring features accounted for a majority portion of the diagnostic variance that is extracted from recorded polygraph data. Table 1 shows the point biserial correlations were reported for information extracted from each recording sensor. EDA data produced an r^2 of .45. When normalized with the other sensor coefficients the relative weight for the EDA was .51.

Nelson, Krapohl & Handler (2008). Described the development of the Objective Scoring System, version 3 (OSS-3), and reported the coefficients from a discriminate analysis. Table 1 shows the normalized structural weighting for the EDA data was .53.

Nelson (2018) described the evolution and development of auto-centering EDA solutions for field polygraph instruments and reported the point-biserial correlation coefficient for of EDA and the criterion state of deception and truth-telling. The r^2 (coefficient of determination) for auto-centered EDA data was .49 and is shown in Table 1. The r^2 for manually centered EDA was .48, suggesting that EDA accounts for approximately half of the variation in deceptive and truthful scores for guilty and innocent examinees.

Nelson [in press] reported the results of a structural weighting function for respiration, EDA, and cardiovascular activity, computed with a simple genetic algorithm. A genetic algorithm is a simple form of machine learning (also known as artificial intelligence) based in the principles of genetics and evolution: random solutions, survival of the fittest, recombination, mutation, and generational

Table 1. Criterion coefficients for respiration, EDA, cardiovascular activity and vasomotor activity.

	Type of coefficient	Non-normalized coefficients				Normalized coefficients			
		Respiration	EDA	Cardio	Vasomotor	Respiration	EDA	Cardio	Vasomotor
Kubis (1962)[1]	df	.18	1.0	.419	-	.12	**.62**	.26	-
Kubis (1964) [1] summed	Prop. correct	.55	.94	.82	-	.24	**.41**	.35	-
Kubis (1964)[1] discriminate	Prop. correct	.56	.85	.76	-	.26	**.39**	.35	-
Kircher (1981) Numerical	df	.28	.65	.37	-	.22	**.50**	.28	-
Kircher (1981) Computer	df	.35	.73	.43	-.27	.20 (.23)	**.41 (.48)**	.24 (.28)	.15
Kircher (1983) Numerical	r_{pb}^2	.32	.37	.28	.36	.24 (.33)	**.28 (.38)**	.21 (.29)	.27
Kircher (1983) Computer	r_{pb}^2	.24	.56	.19	.19	.20 (.24)	**.47 (.57)**	.16 (.19)	.16
Kircher (1983) Table 9	df	.37	.68	.17	.25	.25 (.30)	**.46 (.56)**	.12 (.14)	.17
Kircher & Raskin (1988) Manual	r_{pb}^2	.57	.61	.53	.60	.25	**.26**	.23	.26
Kircher & Raskin (1988) Computer[3]	r_{pb}^2	.30	.59	.37	-	.24	**.47**	.29	-
Raskin, Kircher, Honts & Horowitz (1988)	r_{pb}^2	.15	.53	.48	-	.13	**.46**	.41	-
Raskin & Kircher (1990)	df	-.26	.78	.37	-	.18	**.55**	.26	-
Capps & Ansley (1992)	total scores	2537	3805	3074	-	.27	**.40**	.33	-
Harris & Olsen (1994)[2]	β	-2.6	5.5	3.1	-	.23	**.49**	.23	
Ansley & Krapohl (2000)	frequency	3,455	10,109	4966	-	.19	**.55**	.26	-
Honts Amato & Gordon (2000)	r^2	.12	.39	.17	.26	.13 (.18)	**.41 (.57)**	.18 (.25)	.28
Kircher, Krisjianson, Gardner & Webb (2005)	r_{pb}^2	.18	.45	.26	-	.20	**.51**	.29	-
Nelson, Krapohl & Handler (2008)	df	.629	1.753	.920	-	.19.	**.53**	.28	-
Nelson & Handler (2013)	r_{pb}^2	.19	-	-	-	-	-	-	-
Nelson (2018)	r_{pb}^2	-	.49	-	-	-	-	-	-
Nelson [in press]	proportion	-	-	-	-	.12	**.54**	.34	-

1 These studies used an electronically amplified fingertip plethysmograph instead of the traditional cardio cuff sensor. The fingertip plethysmograph is superseded in contemporary polygraph systems by the photoelectric plethysmograph that is similar to a medical pulse oximeter. Whereas a medical pulse oximeter can acquire and record information on pulse rate, respiration rate and oxygen saturation, the photoelectric plethysmograph in polygraph testing is used to record changes in vasomotor activity.

2 The logistic function also included pulse and pulse line length, though these are not shown in order to simplify the illustration of the relative contribution of EDA, cardiovascular and respiration activity.

3 The multivariate weighting function also included EDA recovery time (duration) and EDA burst frequency (complexity). The discriminate function was: EDA amplitude = .77, EDA recovery time = .27, EDA burst frequency = .28, cardio amplitude = .22 and respiration = -.40.

improvement. Results from optimization with the genetic algorithm are shown in Table 1. When applied to respiration, EDA and cardiovascular data, EDA data accounted for 54% of the diagnostic variance in a sample data of confirmed field polygraphs.

Discussion

This paper is a literature survey of the development of structural weighting coefficients for respiration, EDA, cardiovascular, and vasomotor activity signals used in PDD testing. Nineteen different coefficient functions from 14 different studies are shown in Table 1, along with the results of two additional studies that reported the criterion coefficients for individual sensors. Also shown in Table 1 are the normalized coefficients. Normalized coefficients are a simplification of multivariate structures but offer the advantage of easier and more intuitive comparison of different types of coefficients.

The included studies cover a wide time span, from 1962 to the present. Published studies have employed a variety of methods to evaluate the strength of relationship and structural contributions of scores from different PDD recording sensors and the criterion states of deception and telling. Although there is some variability in different estimates of weighting coefficients there is obvious consistency in that EDA data accounted for greater proportion of diagnostic variance than other sensors for all studies included in Table 1. The mean of all normalized coefficients for the EDA data in Table 1 was .45. when the vasomotor sensor was included in the normalization, and .49 without the vasomotor sensor.

Vasomotor activity sensors have been used inconsistently and are not included in many studies. When it is included, the normalized proportion of the vasomotor sensor has varied from .15 in the numerical scores of Kircher (1981) to .28 in the normalized correlations of Honts, Amato and Gordon (2000). Vasomotor activity was not included in the discriminate function of Kircher and Raskin (1988), though the point-biserial coefficient was reported for the manual scores in this study. This suggests that vasomotor activity may not have contributed additional diagnostic variance to the computer model even though the vasomotor data is correlated with differences between deception and truth-telling in the PDD testing context. Reasons for this are not completely understood and may be incompletely explored. It is reasonable to assume that vasomotor activity would have been included in a computer function if it contributed additional diagnostic variance. It is possible that vasomotor activity may covary strongly with both cardiovascular activity, and EDA, and this may be related to the absence of the vasomotor data in the discriminate function. Further research is needed in this area.

Of the 19 normalized functions, 16 of them produce an EDA weighting coefficient over .4. None of the normalized coefficients for the other sensors exceeded that of the EDA. However, one study, involving the manual scores of Kircher and Raskin (1988) included a vasomotor coefficient that equaled that of the EDA (.26). One other study, involving the manual scores of Kircher (1983) reported a vasomotor coefficient (.27) that nearly equaled that of the EDA (.28).

An obvious limitation of this project is that no attempt was made to test the significance of observed differences between the included studies. Also, no attempt was made to test the difference between the sensor data within the included studies. Another, necessary, caution is in order when attempting to interpret normalized correlation coefficients. This is because correlation coefficients are calculated for individual sensors and does not account for covariance between the sensors. Some of the included studies did report a structural model, and these may offer better information than simple normalization of the reported coefficients.

Finally, it is important to remember that neither EDA, nor any of the other sensor data, are expected to be a perfect, deterministic, indicator of deception and truth-telling. EDA is often discussed in the context of the sweating metaphor. However, just as EDA is not synonymous with deception, EDA is also not synonymous with sweating. Both EDA and sweating are associated with increased activity in the autonomic nervous system, and the array of polygraph recording sensors is almost uniformly autonomic. However, neither

sweating nor autonomic activity are synonymous with or deterministic of deception. That is autonomic activity and sweating can occur for other reasons. Sweating is merely a convenient metaphor for EDA.

EDA data in field polygraph testing is measured using electrical measurements: Ohms or Siemens. However, EDA is not synonymous with electrical resistance or electrical conductance. That is, non-human objects can have electrical resistance and electrical conductance without EDA. EDA is a complex phenomenon for which electrical measurements are a convenient and expedient form of data acquisition. Just as electrical resistance and electrical conductance are a proxy for EDA, EDA itself is a proxy for autonomic activity, while autonomic activity is a proxy for deception. Proxy information is not adequate alone and may be more adequate when combined with other information. Other measurement technologies, besides resistance and conductance, exist for EDA data.

If the normalized coefficients from these studies are interpreted as an indicator of the proportion of test scores and test results that is explained by each recording sensor relative to the other sensors, cardiovascular activity data may account for approximately 30% of observed PDD results, while respiration data may account for approximately 20% of observed results. EDA data may account for approximately 50% of observed PDD scores and test results. Automated computer scoring algorithms have frequently made use of the differences in the contributions of the different recorded PDD signals. Also, differences in structural weight or contribution can be observed in manual PDD scoring methods based on the seven-position Likert-type scale (Likert, 1932). However, manual scoring methods that make use of the unweighted three-position scale may be ignoring some of the diagnostic variation it recorded PDD test data.

One important, and sometimes easily overlooked, difference between seven-position and three-position scales in PDD test data analysis is that the seven-position scale is a Likert-type scale – intended to transform subjective information to numerical values – whereas the three-position scale can be characterized as an objective rank scale. In other words, differences in seven position scale values are subjective or arbitrary (i.e., without mathematical proof as to the selection of differences in scale values) whereas three-position scale values can be subject to objective mathematical proof as to all differences in scale values. Although seven position scores can achieve a similar approximation, seven-position scores are likely to remain less reliable than three-position scores in field practice, due to the lack of theoretical and mathematical proof as to differences in seven-position scale values – leading to either subjectivity or arbitrariness (i.e., arbitrary use of mathematical ratios) in score assignment, or to a non-trivial optimization problem that will require a volume of high quality data and analytic effort. Doubling the EDA scores of three-position manual scores is a simple and objective way to closely and objectively approximate the optimal structural solution that can be achieved through more complex statistical methods.

Weighted three-position EDA scores make use of long-standing knowledge about differences in the structural contribution of different PDD signal, do so in a manner that does not introduce additional subjectivity, arbitrariness and unreliability to the analytic process. It is therefore not surprising that some non-parametric feature extraction and numerical transformation methods, such as the Objective Scoring System (Krapohl, 2002; Krapohl & McManus, 1999) and the ESS/ESS-M (Nelson, Krapohl & Handler, 2008; Nelson et al., 2011; Nelson, 2017), have reported some advantages in manual scoring when weighting the EDA data more than the other sensor data. Although there are some known advantages to data analytic and machine-learning methods that are deliberately naive as to the structural contribution of different signals – especially in the early stages of the development of analytic solution – multivariate solutions that can make use of available knowledge about the relative strength of different signals have ultimately tended to be more powerful or effective. Continued interest is warranted in the differences in correlation and structural contribution of different PDD recording sensors and the potential for optimization and improvement of PDD test effectiveness that may be achieved through the strategic use of naive and weighted structural solutions in PDD analytics.

References

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

Capps, M. H. & Ansley, N. (1992). Numerical scoring of polygraph charts: What examiners really do. *Polygraph, 21,* 264-320.

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

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

Honts, C. R., Amato, S. & Gordon, A. (2000). *Validity of outside-issue questions in the control question test.* Final Report on Grant No. N00014-98-0725.

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

Kircher, J. C. (1983).C*omputerized decision making and patterns of activation in the detection of deception.* [Doctoral dissertation.] University of Utah, Salt Lake City.

Kircher, J. C., Kristjansson, S., Gardner, M. K., & Webb, A. K. (2012). *Human and computer decision making in the psychophysiological detection of deception. Polygraph, 41(2),* 77-126.

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

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

Kubis, J. F. (1962). *Studies in Lie Detection: Computer Feasibility Considerations.* Final report, 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.

Kubis J. F. (1964). *Analysis of Polygraphic Data.* Final report, U.S. Air Force contract RADC-TDR-64-101, AF30(602)-2634. Air Force Systems Command, U.S. Air Force, Griffiss Air Force Base. New York: Rome Air Development Center.

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

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

Nelson, R. (2018). 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. [in press]. Structural Optimization of Respiration, EDA and Cardio Activity Using a Genetic AI. *Polygraph & Forensic Credibility Assessment,* [in press].

Nelson, R., & Handler, M. (2013). Pneumograph signal processing and feature extraction. *Polygraph, 41(3),* 163-172.

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.

Raskin, D. C. & Kircher, J.C. (1990, May 2). *Development of a Computerized Polygraph System and Physiological Measures for Detection of Deception and Countermeasures: A Pilot Study.* Contract 88-L655300-000. Scientific Assessment Technologies, Inc. [Reprinted in Polygraph,30, (3), 153-162.]

Raskin, D. C., Kircher, J. C., Honts, C. R. & Horowitz, S.W. (1988). *A study of the validity of polygraph examinations in criminal investigations*. Final Report, National Institute of Justice, Grant No. 85-IJ-CX-0040. [Reprinted in Polygraph & Forensic Credibility Assessment , 2019, 48 (1)].

Automated Analysis of the Marin Dataset with the ESS-M

Raymond Nelson and Mark Handler

Abstract

The Marin dataset was evaluated with a fully automated version of the ESS-M. Automation was applied to feature extraction, numerical transformation and data reduction, calculation of likelihood statistic for each of the sample cases, and the execution of decision rules to obtain a categorical test result from the numerical and probabilistic test data. The automated ESS-M achieved a decision accuracy rate of 93.5%, excluding an 8.0% inconclusive rate with alpha = .05 for both deception and truth-telling - an accuracy effect commensurate with the APA Standards of Practice requirements for evidentiary exams. Effect sizes and confidence intervals for test sensitivity, specificity, false-positive and false-negative errors are shown. Effects are also shown for other automated scoring methods, including, the OSS-2, OSS-3, and Probability Analysis algorithms. No significant difference was found between the manual and algorithm scores for correct decisions or inconclusive rates.

Introduction

Analysis of psychophysiological detection of deception (PDD) test data is a structured process involving a coherent sequence of operations that is fundamentally similar to many other scientific tests. Analytic procedures include feature extraction, numerical transformation and data reduction, calculation of a likelihood statistic for the observed test data, and the use of a structured rule to parse a categorical test result from the numerical and probabilistic information.

A traditional approach to PDD test data analysis is to execute the analytic process manually, using visual feature extraction, limited transformations involving only simple addition, numerical cutscores that can be memorized for use without reference to statistical formulae or reference tables, and procedural decision rules that can be reduced to intuitive heuristics.

When reduced to rules and procedures, manual scoring methods can be learned, practiced and executed with some reliability by human experts. Human PDD experts using visual and manual PDD scoring methods are, however, potentially capable of exhibiting inconsistency and imperfect reliability. Computer algorithms, on the other hand, are intended to execute these same operations quickly and with automated reliability. This study is a demonstration of the capabilities of an automated Empirical Scoring System – Multinomial (ESS/ESS-M Nelson, 2017a; 2017b) analysis algorithm, with automated feature extraction capabilities that were designed to be robust against difficult data. The automated ESS-M algorithm executed scoring tasks using the same ESS/ESS-M transformations (Nelson et al., 2011) and Bayesian calculations (Nelson, 2018a; Nelson & Rider, 2018; Nelson, Handler, Coffey & Prado, 2019) that are used by human experts when manually scoring PDD data.

Method

The sample data for this study came from an archival dataset consisting of 50 confirmed guilty exams and 50 confirmed innocent exams that were conducted using the

The authors are grateful to Don Krapohl for his reviews, edits and comments to earlier versions of this manuscript.

U.S. Federal Zone Comparison Test (ZCT) format (Department of Defense, 2006). The Federal ZCT consists of three relevant questions (RQs) and three comparison questions (CQs), along with other procedural questions that are not subject to analysis. Although some differences in inconclusive rates have been shown to occur as a function of the number of RQs, no significant differences in criterion accuracy have been shown for the variety of test formats that include three RQs and three CQs (American Polygraph Association, 2011). The Federal ZCT format is structurally and substantively similar to other test formats that include three RQs and three CQs.

This sample has been used in previous studies, and is referred to as the Marin dataset, because it was selected for use when evaluating examiners for the Marin protocol (Krapohl, 2005; Krapohl & Cushman, 2006, Marin, 2000). The Marin dataset is useful for this project for several reasons, including the fact that evaluation of the cases has been described as a challenging, though manageable, task for human experts, and the fact that published accuracy effect sizes are available for both human scorers and earlier algorithm projects (Nelson, Krapohl & Handler, 2008).

The Marin dataset was selected randomly from the 2002 confirmed case archive at the U.S. Department of Defense, with the constraint that the sample will include a balanced number of confirmed guilty and confirmed innocent cases. Examinations in the Marin dataset were conducted during the 1990s by a variety of federal and civilian law enforcement agencies and were confirmed through a combination of confession and extra-polygraphic evidence. However, it is not known how the cases came to be included in the confirmed case archive.

The Marin sample data are interesting for several reasons, including that it has been used in previous studies that provide opportunity for comparison of effect sizes for different analysis methods. Table 1 shows previously reported accuracy effects when using the Marin dataset with manual scores, including result from experienced examiners in Krapohl (2005) and Krapohl and Cushman (2006) and both experienced and inexperienced examiners in Nelson, Krapohl and Handler (2008). Result shown are means of the human scorers.

Table 1. Previously reported results (point estimates) using the Marin sample with manually scored results.

	Percent Correct	Percent Inconclusive	Sensitivity (TP)	Specificity (TN)	FN Errors	FP Errors
Krapohl (2005) [7 position, traditional rules]	.896	.221	.783	.615	.068	.093
Krapohl (2005) [7 position, evidentiary rules]	.896	.096	.808	.803	.104	.089
Krapohl & Cushman (2006) [7 position, traditional rules]	.861	.198	.824	.556	.044	.132
Krapohl & Cushman (2006) [7 position, investigative rules]	.872	.073	.792	.824	.122	.086
Nelson, Krapohl & Handler (2008) [ESS, .05/.10]	.875	.102	.776	.802	.129	.096

Analysis

Data for all cases were exported from a proprietary digital format to the NCCA ASCII format, and then imported to the R language and environment for statistical computing for all analysis (R Core Team, 2019). ESS/ESS-M feature extraction (Nelson *et al.,* 2011) for each sensor, for each relevant question (RQ), and each comparison question (CQ) was completed automatically using the Kircher features (Krapohl & McManus, 1999; Kircher, Kristjiansson, Gardner, & Webb, 2005). All ESS/ESS-M numerical transformations were also completed via automated computer algorithm, including the selection of a single CQ for each RQ. Automation was also used to complete all statistical lookups and calculations necessary to obtain the ESS-M likelihood statistics for the sample cases, and to compute the Bayesian posterior likelihood of deception or truthtelling (Nelson, 2018a; 2018c; 2018d). Finally, execution of two-stage decision rules, aka, Senter Rules, (Senter, 2003; Senter & Dollins, 2003) was completed via automation.

Feature extraction

Respiration data were measured as the mean of respiration line excursion (RLE; the absolute difference of each subsequent respiration sample) (Kircher & Raskin, 2002) for a one-second moving average from stimulus onset to 15 seconds post stimulus onset excluding the data from one second before to one second after the recorded verbal answer. This measurement is robust against distortions at the point of verbal answer and is not influenced by the length of the 15 second evaluation window – effects with different measurement periods will have a similar metric. EDA reactions were measured as the onset of a positive slope segment during a response onset window (ROW) from .5 seconds after stimulus onset to 5 seconds after the verbal answer to the greatest y-axis (vertical) distance to subsequent peak of reaction (onset of negative slope) within an evaluation window from stimulus onset to 15 seconds after stimulus onset. Cardiovascular activity was extracted by first calculating the mean of all cardio sensor samples. This can be thought of, and plotted, as the mid-line between the systolic and diastolic peaks. Cardiovascular activity changes were then extracted, using the cardio mid-line, using a procedure similar to the one for the EDA data.

One additional procedure was included in the automated feature extraction for the EDA and cardiovascular data. If there was no response onset during the ROW, a response onset was inferred statistically during positive slope segments using a *z*-test of the variance of the one second mean of the difference of each subsequent EDA sample. A response onset was imputed if the difference in variance for two adjacent one-second windows exceeded the alpha = .001 boundary. This can be visualized as a substantial increase in positive slope angle within a positive slope segment during the ROW. All measurement values were taken in dimensionless units – not indexed to any physical quantity.

Numerical transformation and data reduction

For each recording sensor, dimensionless values for all RQ and CQ measurements were transformed to R/C ratios by pairing each RQ with a single CQ using the heuristic described by Nelson (2017c). For this test format the first RQ in the question sequence is compared to the preceding or subsequent CQ depending on which CQ has produced the greater change in physiology, while the second and third RQs are compared only to the preceding CQ. Each of the R/C ratios were then transformed to their natural logarithm. Use of the log transformation changes the asymmetrical distribution of all possible ratios (between zero and infinity, with a mean of one), to a symmetrical distribution with a mean of zero.

Sign values of respiration data were inverted so that all logged R/C ratios are intuitively similar to traditional integer scores used when manually scoring polygraph data. In this way, negative scores are associated with deceptive test outcomes and positive scores are associated with truthful outcomes. Finally, all logged R/C ratios were converted to three-position objective rank scores (Department of Defense, 2006; Nelson & Handler, 2018) using the ordinal rank values [-1, 0, +1] using the optimization constraints shown in Table 2.

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. Previous optimization studies indicated that + scores are more likely to be negatively correlated with truth-telling without the asymmetrical constraint.
** The upper limit constraint for electrodermal and cardiovascular activity served only to reduce the likelihood that a numerical score is extracted from an instability artifact.

Logged R/C ratios greater than zero (positive values) were transformed to -1 integer scores if they were within the optimization constraint values for positive scores, while logged ratios less than zero (negative values) were converted to +1 integer scores if they were within the optimization constraints for negative scores. Values that were not within the minimum and maximum optimization constraints for negative or positive scores were transformed to ordinal rank scores of zero (0). All sign values were corrected so that positive (+) scores correspond to truthful outcomes while negative (–) scores correspond to deceptive outcomes.

Three-position ordinal rank scores were then transformed to ESS/ESS-M scores by doubling all EDA scores (Nelson, 2017a; Nelson *et al.,* 2011). Previous studies have shown that EDA data have a stronger criterion correlation and greater structural contribution polygraph effect sizes than other sensor data, often contributing nearly one-half of the information contained in the final score and test result [See Nelson (in press) for a summary].

ESS/ESS-M scores were then summed between recording sensors to subtotal scores for each individual RQ. Subtotal scores were summed to obtain a grand total score. The resulting ESS/ESS-M scores can be viewed as analogous to the ESS/ESS-M values that would be obtained via traditional manual/visual feature extraction – with the advantage that automated reliability provides greater resistance to human error due to fatigue, confusion or other bias . Like traditional manual scores, automated ESS-M scores are expected to be loaded in the range greater than zero for innocent examinees, and less than zero for guilty examinees.

ESS-M likelihood function and Bayesian calculations.

The purpose of any scientific test is to quantify some phenomena that cannot be subject to perfect deterministic observation or direct physical measurement. Virtually all scientific tests require the use of some form of likelihood function to provide a statistical or probabilistic value for the observed test data [See Casella and Berger (2003) for information on statistical inference.] A likelihood function can be as simple as the known test sensitivity or specificity rates. A likelihood function can also take the form of a mathematical equation or statistical function – or can also take the practical form of a computer program to accomplish the mathematical and statistical calculations. A likelihood function can be found in other forms such as an empirical reference distribution. And a likelihood function may even be calculated from a theoretical reference

1 A difference between automated ESS/ESS-M scores and manual ESS/ESS-M scores – aside from the difference in process – is that automated scores are rank scores whereas manual scores can be thought of as Likert (1932) scores. Likert scores are a highly useful device to code or transform subjective information to numerical values for analysis. In this context the difference between rank scores and Likert scores is that rank scores are objective, leading to improved reliability and improved ability to quantify the margin of error or level of confidence associated with a test result or conclusion.

distribution – using only information subject to mathematical and logical proof under the theory of a test.

The ESS-M likelihood function is this latter type – a multinomial distribution (reference table) of all possible PDD test scores, under the null-hypothesis to the analytic theory of PDD testing, and the statistical likelihood of each possible score. The analytic theory of PDD testing holds that greater changes in physiological activity are loaded on different types of stimuli as a function of deception or truth-telling in response to relevant target stimuli (Nelson, 2015; 2016). The null hypothesis to the analytic theory of the polygraph states that PDD scores are not loaded in any systematic way.

In many areas of science and scientific testing, it is very difficult to calculate an expected distribution under a theory or hypothesis, but it is quite easy to calculate a distribution under a null hypothesis. This is because the null hypothesis can often be characterized as resulting in a random distribution of scores. The expected distribution of randomly loaded ESS/ESS-M scores under the null hypothesis is multinomial (Nelson, 2017a). The multinomial distribution of ESS/ESS-M scores can provide a mathematically coherent and reproducible likelihood value for an observed test score. The likelihood value can be submitted, together with the observed test data and a prior probability of deception or truth-telling, to Bayes theorem (Bayes & Price, 1763; Berger, 1985; Laplace, 1812; Rubin, Gelman, Carlin. & Stern, 2003; Stone, 2013, Winkler, 1972) to calculate a posterior probability of deception or truth-telling.

For each case, a Bayesian confidence interval (credible interval) was calculated for the ESS-M posterior probability of deception or truth-telling using the Clopper-Pearson method (Clopper & Pearson, 1934) and a one-tailed alpha = .05th for deception and truth-telling. In this way, cases were classified as deceptive when the .05th percentile lower limit of the posterior odds of deception exceeded the prior odds of one-to-one. Similarly, cases were classified as truthful when the .05th percentile lower limit of the posterior odds of truth-telling exceeded that same prior. Cases were classified as inconclusive when the prior odds (1 to 1) were within the Clopper-Pearson interval. To field polygraph examiners this can be simplified using lookup tables (Nelson, 2017a; 2018a) which show the ESS-M numerical cut-scores for grand total scores were +3 or greater for truthful classifications, and -3 or lower for deceptive classifications. An additional attempt was made to classify the cases using the subtotal scores when the grand total score was inconclusive. This was done using a subtotal cut-score of -7 or lower for the statistically corrected posterior odds of the lowest of the three RQ subtotal scores.

Decision rules

All scientific test results are fundamentally probabilistic, simply because the purpose of many scientific tests is to quantify phenomena of interest that are amorphous and not subject to physical measurement. Probabilistic results are interesting for scientific purposes, but often provide inadequate structural guidance for practical purposes. Scientific test results, in practical terms, are interpreted categorically as either *positive or negative*, wherein the term positive signifies the presence of the phenomena of interest while negative signifies the absence of the phenomena of interest. Categorical test results for PDD exams are often reported using the terms deception indicated (*DI*)or no deception indicated (*NDI*) for diagnostic exams and significant reactions (*SR*) and no significant reactions (*NSR*) for screening examinations.

Regardless of the terminology used, the meaning of the categorical test result is unchanged. Of importance is the procedure employed to parse the categorical result from the numerical and probabilistic information. A number of PDD decision rules are described in publication. Categorical results for Federal ZCT exams in the Marin dataset were parsed automatically using a two-stage decision rule (Senter, 2003; Senter & Dollins, 2003). The Two Stage Rule (TSR) is executed in two stages, the first of which involves the aggregation of all scores for all RQs. The second stage of the TSR is employed only if the results of the first stage are inconclusive and involves the use of the individual RQ scores to attempt a positive classification of the test result. In practice only one RQ score, that most indica-

tive of deception, is used during the second stage. [See Nelson 2018b for a survey and description of PDD decision rules.]

Results from the automated ESS-M were tabulated for each case. Results were then sent to a third party for comparison with the known criterion states for each of the sample cases. Information returned from the analysis included the TP, TN, FN, FP, and inconclusive rates, but did not include the criterion states of individual cases. As a result, no post-hoc investigation of errors was possible. Confidence intervals were calculated in the R language and environment for statistical computing (R Core Team, 2019) using a bootstrap procedure.

Results

Results of the automated ESS-M with the Marin dataset are shown in Table 3, including the point estimate for the sample dataset and bootstrap 95% confidence interval (.025th and .975th quantiles) for the percent correct, rate, along with the sensitivity or true-positive (TP), specificity or true-negative (TN), false-positive (FP) and false-negative (FN) rates. The automated ESS-M achieved a decision accuracy rate of 93.5%, excluding an 8.0% inconclusive rate. To facilitate a more direct comparison of the automated ESS-M and OSS-3 algorithms, results were recalculated with the OSS-3 (Nelson et al., 2008) using a symmetrical alpha boundary of .05 for deceptive classifications and .05 for truthful classifications. These results were only slightly different from the previously reported OSS-3 results (where alpha was .05 for deception and .1 for truth-telling), with a decision accuracy rate of 91.9% excluding a 14.0% inconclusive rate.

Table 3. Point estimates and confidence intervals [.025th, .975th} for automated ESS-M and OSS-3 with the Marin dataset (alpha = .05, .05).

	Percent Correct	Percent Inconclusive	Sensitivity (TP)	Specificity (TN)	FN Errors	FP Errors
Automated ESS-M [alpha = .05,.05]	.935 {.868, .969}	.080 {.030, .120}	.920 {.837, .981}	.800 {.683, .905}	.040 {.001, .104}	.100 {.021, .189}
OSS-3 [alpha = .05,.05]	.919 {.855, .968}	.140 {.080, .210}	.860 {.756, .948}	.720 {.591, .840}	.040 {.001, .102}	.100 {.021, .191}
Probability Analysis [probability cutoffs = .3, .7]	.905 {.837, .964}	.150 {.080, .220}	.778 {.654, .886}	.761 {.635, .875}	.122 {.039, .220}	.040 {.001, .104}
OSS-2 [cut-scores +8,-8]	.925 {.857, .974}	.200 {.080, .210}	.760 {.659, .887}	.720 {.683, .905}	.080 {.021, .191}	.040 {.001, .104}

Also shown in Table 3 are the effect sizes and bootstrap confidence intervals for confidence for the Probability Analysis (PA) algorithm (Kircher and Raskin, 1988; Raskin, Kircher, Honts & Horowitz, 1988) along with the OSS-2 algorithm (Krapohl & McManus, 1999; Krapohl, 2002) using the [+8, -8] cut-scores recommended by Dutton (2000), and the OSS-3 algorithm (Nelson, Krapohl & Handler, 2008).

A four-level one-way ANOVA for the percentage of correct decisions was not significant [F (3,396) = .179, (p = .910)] for differences among the automated scoring algorithms. A second for level ANOVA for inconclusive results was also not significant [F (3,396) = 2.253, (p = .082)]. Post-hoc power analysis showed the experimental power = .072 to be weak for the detection of a significant accuracy effect size difference similar to that observed and would require a sample size of N=3,030 to achieve a power level of .8. Post-hoc power calculations for the observed difference in inconclusive rates also showed the experimental power = .496 to be weak at detecting a significant difference in effect size as small as that observed, with a required sample size of N=191 to achieve a power level of .8. Mean decision accuracy for the four automated scoring algorithms was 92.1% and the mean inconclusive rate was 14.3%.

A second three-level one-way ANOVA was not statistically significant for differences in decision accuracy for the manual scores – using only the results with evidentiary rules – shown in Table 1 [F (2,297) = .150, (p = .861)]. Differences were also not significant for inconclusive results for the manual scores [F (2,297) = .289, (p = .749)]. These results replicate the results of earlier studies (Horvath, 1974, Nelson, Krapohl & Handler, 2008; Raskin, Kircher, Honts & Horowitz, 1988). that showed no effect for the level of experience when analyzing polygraph data

A bootstrap hypothesis test of the n=100 sample case results was used to compare accuracy for the mean of manual scores and the mean of automated scores. Although the mean accuracy for the computer algorithms exceeded the mean accuracy for manual scores, results from the bootstrap hypothesis test were not statistically significant (p = .266). The difference in the observed inconclusive rate for the computer algorithms and manual scores was also not statistically significant (p = .111). As described earlier, a sample of substantially larger size would be required to have sufficient statistical power to reliably detect a significant effect of the size of the observed difference between the manual and automated scores in this project.

Discussion

This project is a demonstration of PDD accuracy effect sizes for the Marin dataset, consisting of 100 field exams conducted with the Federal ZCT format, using automated feature extraction of ESS scores with multinomial reference distribution and Bayesian analysis. All aspects of the test data analysis were automated, including feature extraction, numerical transformation and data reduction, calculation of likelihood statistic for each of the sample cases, and the execution of decision rules to obtain a categorical test result from the numerical and probabilistic test data. The automated ESS-M achieved a decision accuracy rate of 93.5%, excluding an 8.0% inconclusive rate. Noteworthy in Table 3 is the fact that each of the four automated scoring algorithms achieved an accuracy rate commensurate with the 90% accuracy requirements of the APA Standards of Practice for evidentiary exams (American Polygraph Association, 2011; American Polygraph Association, 2018). However, the means of manually scored results, shown in Table 1, did not achieve the evidentiary standard, though they were similar to the algorithm results when scoring the same data set.

An obvious advantage of automated analysis is the reliability and reproducibility of analytic results. Of course, a limitation of automated systems is the temptation for field practitioners to attempt to evaluate data that are artifacted, unstable, or unresponsive quality or examination data that is inconsistent with the algorithm's design and intended data requirements. Future development efforts should be devoted to the automation of the evaluation of standards compliance and the interpretable or usable quality of examination data.

One potential advantage of visual feature extraction and manual scoring – related to subjectivity and reliability – is the ability to creatively resolve problems when there are minor departures from standard procedure in the conduct and recording of an exam (e.g., questions labeled in unexpected ways) and when the examination data are unresponsive, artifacted or unstable. This potential advantage also leads to potential problems with reliability of analytic results. Although some human experts may be highly skilled at working with difficult data, it is inevitable that different human scorers may exhibit differences in terms of abilities, levels of fatigue, biases and other motivations. It is also inevitable that any examination that is intended for use in an adversarial legal proceeding – or any adjudication process that involves a decision about a person's rights, liberties or economic and professional opportunity – may be discussed by different experts with different views and conclusions. Additionally, examiners who work with a high volume of examinations should not be surprised (and should probably expect) to observe occasional differences in their own scoring of the same data sets. These concerns have the potential to result in possible confusion and mistrust around the scientific value and meaning of PDD test results. Although computerized PDD scoring algorithms have not possessed the innate creativity or ethically derived motivation of experienced human ex-

perts, computer algorithms available today can "learn" to execute complex procedures with such great speed and reliability that microprocessors, computer logic, and computer algorithms are used to improve the effectiveness of the human professional in virtually every area of science, technology and professional activity.

Analysis of PDD test data has progressed through several stages of development. The earliest involved the development of technology to obtain recordable physiological signals that are correlated with differences between deception and truth-telling. In professional practice, the earliest stages of activity can be thought of as a stage of experimentation wherein creative observers dedicated countless hours to develop an understanding and intuition about complex and high-dimensional data. A subsequent stage of professional activity can be thought of as a stage of expertise in which professional effectiveness and observed effect sizes are a function of expertise - often developed through a combination of innate talent and ability and exhaustive practical trial and error experience under the supervision of more other experts. This was characterized by an emphasis on unstructured professional judgment due to the absence of well-defined structural methods. *Evidence-based practice* is characterized by a well-developed understanding of learnable and reproducible professional activities that are shown to be empirically correlated with desired outcomes.

Clearly defined evidence-based practices will also give rise to the potential for automation of routine and repetitive tasks. Automation increases the potential for reliable execution of complex procedures. Although capable of effect sizes similar to computerized statistical algorithms, visual feature extraction and manual test data analysis are inherently more subjective than automated processes. Also, manually executed procedures can be less reliable than automated algorithms due to human factors such as fatigue, expectation bias and other factors. Although continued interest in both manual and automated test data analysis solutions is recommended, a challenge for professionals will be to learn to make use of automation, autonomous systems and automated analytic methods without neglecting the ethical locus of responsibility that all decisions that affect human outcomes will remain a human concern.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph, 40,* 194-305. [Electronic version] Retrieved August 20, 2012, from http://www.polygraph. org/section/research-standards-apa-publications.

American Polygraph Association (2018). APA Standards of Practice(Effective September 1, 2018). Retrieved (January 20, 2019) from https://www.polygraph.org/apa-bylaws-and-standards.

Bayes, T. & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53,* 370–418.

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

Casella, G. & Berger, R. (2002). *Statistical Inference.* Pacific Grove: Duxbury.

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.

Dutton, D. (2000). Guide for performing the objective scoring system. *Polygraph, 29,* 177-184.

Horvath, F. S. (1974). *The accuracy and reliability of police polygraphic (lie detector) examiners' judgments of truth and deception: The effect of selected variables.* [Doctoral Dissertation.] Michigan State University.

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

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

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

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

Krapohl, D. J. & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: A replication. *Polygraph, 35(1),* 55-63.

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

Laplace, S. P. (1812). *Théorie analytique des probabilités.* Paris: Courier.

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

Marin, J. (2000). He said/ She said: Polygraph evidence in court. *Polygraph, 29,* 299-304.

Nelson, R. (2015). Scientific basis for polygraph testing. *Polygraph 41(1),* 21-61.

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

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

Nelson, R. (2017b). Updated numerical distributions for the Empirical Scoring System: An accuracy demonstration with archival datasets with and without the vasomotor sensor. *Polygraph & Forensic Credibility Assessment, 46 (2),* 116-131.

Nelson, R (2017c). Heuristic principles to select comparison and relevant question pairs when scoring any CQT format. *APA Magazine, 50(1),* 73-83.

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

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

Nelson, R. (2018c). Practical polygraph: a tutorial (with graphics) on posterior results and credible intervals using the ESS-M Bayesian classifier. *APA Magazine, 51(4),* 66-87.

Nelson, R. (2018d). Five minute science lesson: Bayes' theorem and Bayesian analysis. *APA Magazine, 51(5),* 65-78.

Nelson, R. (in press). Literature survey of structural weighting of polygraph signals: why double the EDA? *Polygraph and Credibility Assessment,* [in press].

Nelson, R. & Handler, M. (2018). Reducing Inconclusive Results: A descriptive analysis of decision rules, weighted electrodermal scores and multinomial cut-scores. *Polygraph and Credibility Assessment,* 47(2), 108-121.

Nelson, R., Handler, M., Coffee, T. & Prado, R. (2019). How to: a step-by-step worksheet for the Multinomial-ESS. *Polygraph & Forensic Credibility Assessment , 48 (1),* 60-75.

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.

Nelson, R. & Rider, J. (2018). Practical polygraph: ESS-M made simple. *APA Magazine, 51(6),* 55-62.

R Core Team (2019). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raskin, D.C., Kircher, J.C., Honts, C.R. and Horowitz, S.W. (1988) A Study of Validity of Polygraph Examinations in Criminal Investigations. Grant number 85-IJ-CX-0040. Salt Lake City: Department of Psychology, University of Utah. [Reprinted in Polygraph and Credibility Assessment, 48(1), 10-39.

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

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

Senter, S. M. & Dollins, A. B. (2002). *New Decision Rule Development: Exploration of a two-stage approach.* Report number DoDPI00-R-0001. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC. Reprinted in Polygraph 37, 149-164.

Stone, J. (2013). *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis.* Sebtel Press.

Winkler, R. L. (1972). *An Introduction to Bayesian Inference and Decision.* Holt Mc-Dougal.

Editor's note: update to the National Center for Credibility Assessment PDD 503-ANALYSIS II TEST DATA ANALYSIS: Numerical Evaluation Scoring System Pamphlet.

There is an updated version of this manual dated August 2017, available for download at:

https://antipolygraph.org/documents/ncca-numerical-scoring-2017-08.pdf

Here is the note describing the changes.

> Note: The content and substance of this pamphlet was officially approved on 23 Aug 2006. In March 2011, this pamphlet was edited, but only for grammatical issues and curriculum style, format, and consistency of design factors, and no substantive changes to the content were made. In January 2012, content changes were made relative to requirements for two artifact free askings. In June, 2014, content changes were made effectively eliminating the 2:1 ratio from the EDA Ratio Scale. In Aug 2015, content changes were made changing the wording of "artifact free" to "valid asking", as well as a definition added to App. G. Jan 2017: Changed EDA/CV ROW from "SO to EA" to "SO to 5 seconds beyond the EA". The latency rule is effectively eliminated by the new ROW. Aug 2017: Updated the description of the 7-position scoring methods for the CV channel.

Introduction to the NCCA ASCII Standard

[editorial staff]

The NCCA ASCII format is a specification that defines a structured file format that all North American polygraph instrument manufactures were requested to include in all software versions beginning during 2009. The NCCA ASCII text format offers the potential for readability by human or machine, and resolves data access problems for research, development and analysis in contexts that may involve different polygraph instrument software solutions with different proprietary formats. The capability to export data to a common format reduces the likelihood that vendor formats will be become known or compromised. A common format also increases the capability to accommodate future changes to proprietary data formats and reduces the likelihood that valuable data will become obsolete or unusable. An ability to access polygraph data in a known data output specification improves options for polygraphic feature extraction and data analysis beyond the visual and manual methods that were the best available technology solutions during the pre-computer epoch. The specification and an example are shown in Appendices.

During 2009 the Threat Assessment and Strategic Support (TASS) branch of the Defense Academy for Credibility Assessment (DACA; now known as the National Center for Credibility Assessment, NCCA) requested all North American manufacturers of polygraph instruments to include a software feature to export the recorded polygraph data to a known ASCII text format, referred to here as the NCCA ASCII specification. The goal was to create a data format that contains all the hardware, software, physiological data, timing, and question information for each chart in an examination.

The specification itself was defined by Dr. Andrew Dollins [now deceased] and Dr. John Kircher. In the 10 years since the definition and implementation of the NCCA ASCII specification, its importance and usefulness have been subtle for most field polygraph examiners but substantial for those interested in accessing polygraph data for scientific study or research. Whereas most field examiners view the displayed or printed chart tracings (a term held over from the days when ink was traced onto a moving paper via capillary action) as the data itself, scientists interested in studying or advancing our knowledge and field practices in scientific lie detection will view data as a time-series of recorded numerical values – for which computers are unquestionably useful.

Today nearly all computerized polygraph systems include a convenient tool to export the recorded data to the NCCA ASCII format. A well-specified and known data format person permits researchers and developers to import the data into any preferred statistical or analysis environment. Access to the recorded signals permits the study and optimization of signal processing and feature extraction solutions that go well beyond merely drawing a time-series plot for visual inspection, and potentially well beyond the limitations of traditional visual analysis. In addition to data export capabilities, most computerized polygraph systems available today also include the capability to import the data from the NCCA ASCII format – increasing the potential for meaningful evaluation and comparison of similarities and differences in data display and analysis tools available in different software solutions available to polygraph field practitioners. Although perhaps of limited usefulness to many

[1]The APA editorial staff is grateful to John Kircher and the late Andy Dollins for their work in creating and coding the NCCA ASCII format. We are also indebted to APA Past President Raymond Nelson for encouraging and helping create this document.

field practitioners, these capabilities are of vital importance to the advancement and future of the polygraph profession.

Appendix A shows the general specification for the dimensions and formatting of the output file, in addition to the naming of the output file. Structured file names clearly identify the series and charts for each examination. All information for each test chart is output to a single file. Header information (Appendix B) describes the instrumentation, test date, and other information about the exam such as the data sampling rate, stimulus events, and event timings. Standards are also provided for how the physiological data is to be structured for output (Appendix C), including data for each of the recording sensors. An example of an NCCA ASCII output file is shown in Appendix D.

The NCCA ASCII text format offers the potential for readability by human or machine, and resolves data access problems for research, development and analysis in contexts that may involve different polygraph instruments with different proprietary formats. The capability to export data to a common format reduces the likelihood that vendor formats will be become known or compromised. A common output format also increases the capability to accommodate future changes to proprietary data formats and reduces the likelihood that valuable data will become obsolete or unusable. Most importantly, a common and published specification for data output underscores the fact that providing plotted data on a computer display or printed paper is not the same as proving access to the data. Access to the data, for any scientific purpose, requires access to knowledge about the specifications of the data output format. Technology vendors that do not provide easy access to output data in a known and usable format should be requested to prioritize the future and advancement of the polygraph profession by providing access to the polygraph data through the NCCA ASCII specification. Vendors who do not make data available through the NCCA ASCII specification – or other published format – are limiting the future of the polygraph profession by limiting polygraphic feature extraction and data analysis methods to visual and manual calculation methods that were the best available technology solutions prior to the widespread availability of powerful computing platforms.

Appendix A. Output File Structure and Naming Convention

The objective is that all of the hardware, software, physiological data, timing, and question information for one chart shall be contained in one human readable ASCII text file. The ASCII file shall not contain tab characters because tab-character spacing varies among data-viewing programs. The data columns shall be right justified as depicted below.

The specific ASCII text file output will depend on the originating polygraph system, but we ask that the following conventions be followed.

- Use the upper case letter D as the first letter in the file name.
- Use $, &, %, or # as the second character in the file name as follows:

o $ = Axciton

o & = Lafayette

o % = Limestone

o # = Stoelting

- Use a "-" as the third character in the file name.
- The fourth through Last-2 characters will be the originating examination name (see examples below).
- Use a "-" as the Last-1 character in the file name.
- Use the series/ examination number (in hex) as the last character in the file name follows:

o 1 = Series/ Examination 1

o 2 = Series/ Examination 2

o . . .

o 9 = Series/ Examination 9

o A = Series/ Examination 10

o B = Series/ Examination 11

o F = Series/ Examination 15

o NOTE: Axciton & Stoelting-DOS only – Because Axciton and Stoelting-DOS do not differentiate examination types in the file name structure, please always use an X as the Last-1 character.

- The first and second characters in the file name extension should indicate the chart number such that 01 = chart 1, 02 = chart 2, etc.

• Use the letter “A” as the final character in the file name extension.

• Output file name examples:

o D$-$$$$ZG14-X.03A = Axciton, Exam X, Chart 3

o D$-$$765TGZ-X.17A = Axciton, Exam X, Chart 17

o D&-175-802-20050106 MGQT-2.01A = Lafayette, Exam 2, Chart 1

o D&-PF71-2004-03-19-PR1-B ZCT-1.04A = Lafayette, Exam 1, Chart 4

o D%-2007-13-1.01A = Limestone, Exam 1, Chart 1

o D%-DACA 04 2007-3.02A = Limestone, Exam 3, Chart 2

o D#-00001-X.02A = Stoelting, Exam X, Chart 2

o D#-01-X.09A = Stoelting, Exam X, Chart 9

Appendix B. Output ASCII Text File Header Information

• The structure of the ASCII Text output file shall be as follows (see examples at the end of this document):

o Name of the file being written

o Name of the source data file (for Lafayette, Limestone, & Stoelting Windows use the primary subdirectory name)

• Name of the instrument used to collect the data as follows:

o Axciton DOS

o Axciton Windows

o Lafayette Windows

o Limestone Windows

o Stoelting DOS

o Stoelting Windows

• Version of the software used to collect the data, or "No Version Available"

• Date the data was collected in DDMMMYY format

• Time data collection began in military format (1000, 1300, 2000 etc)

• Examination number (or "X" for Axciton & Stoelting)

• Number of the chart in the series:

• Number of questions:

• Number of data samples per second (fastest channel)

• Number of data samples per fastest channel

• Number of channels

• Sample rates (in Hertz) for each channel as follows (assuming variable storage rates):

o Sample Rate (Hz)- UPneumo: 10

o Sample Rate (Hz)- LPneumo: 10

o Sample Rate (Hz)- EDA1: 15

o Sample Rate (Hz)- Cardio1: 60

o Sample Rate (Hz)- Move1: 60

• Event list heading (see examples below)

• Event list as follows: o Columns 01-02 = Event number (01 to 99) Right Justified

o Columns 03-06 = spaces

o Columns 07-14 = Event/ Question Label

o Column 15 = space

o Column 16-80 = Question or Event Text*

* Allowed 8 spaces for the Event/ Question label. This should allow enough space for all vendors to use the Examiner's original Event or Question Label from the original chart – except for Limestone which may have to truncate Question Labels longer than 8 spaces.

** If the Question field is left blank, duplicate the event label in the Question field.

*** If the question text is more than 66 characters long, continue the question beginning in column 16 of the next line.

• Event location heading (See example below)

• Event locations, right justified, as follows: o Columns 01-02 = Event number (01 to 99 0

right justified)

o Columns 03-06 = spaces

o Columns 07-14 – Event/ Question Label (right justified)

o Columns 15-25 = Sample event began on (right justified integer)

o Columns 26-36 = Sample event ended on (right justified integer)

o Columns 37-47 = Sample answer occurred on (right justified integer)

Appendix C. Physiological Data

• The physiological data shall be output into the following eight right justified columns, separated by spaces (if no movement sensor was used the Move1 column shall be filled with the value 9999.9):

- o Columns 01-06 = Data sample number (beginning at 1)
- o Columns 07 = space
- o Columns 08-15 = data sample time (beginning at 0) where XX:YY.ZZ where X=minutes, Y=seconds, & Z= 1/100 of a second
- o Columns 16 space
- o Columns 17-24 = Event/ Question labels right justified with leading "-"s (accurate to fastest sample rate)
- o Columns 25-35 = upper respiration channel data (Upneumo)
- o Columns 36-46 = lower respiration channel data (Lpneumo)
- o Columns 47-57 = electrodermal channel data (EDA1)
- o Columns 58-68 = cardiovascular channel (Cardio1)
- o Columns 69-79 = movement channel (Move1)
- o Columns 80-90 = additional channel 6
- o Columns 091-101 = additional channel 7
- o Columns 102-112 = additional channel 8
- o Columns 113-123 = additional channel 9
- o Columns 124-134 = additional channel 10

• Data columns 1 – 8 shall have the following 8 headings:

- o Sample
- o Time
- o Label
- o Upneumo
- o Lpneumo
- o EDA1
- o Cardio1
- o Move1

• Use the following headings for additional channel data columns as needed:

o Cardio2, Cardio3, - additional ascultatory cardiovascular measures

o PPG1, PPG2 – cardiovascular activity by photoplethysmography

o ECG1, ECG2 – cardiovascular activity by electrocardiography

o Voice1, Voice2 – audio recording channels

o Move2, Move3 – additional movement sensors

o EMG1, EMG2 – electromyographical activity

o EDA2, EDA3 – additional electrodermal activity

o PLE1, PLE2 – plethysmography at other sites (chest, penis)

o TEMP1, TEMP2 – temperature sensors

o LPupD, RPupD – left & right pupil diameter

o LEyeP, REyeP – left & right eye position

o HEOG, VEOG – horizontal & vertical electrooculogram

o Other?

• The data shall be written at the actual sample rate. In EXAMPLE 1 below, each channel was sampled, or interpolated, to 60 samples per second.

NOTE: Limit sample rate to a maximum of 120. To the best available know-ledge the fastest current storage rate is 100 Hz.

• The physiological data shall be written right justified, with 1 decimal place (F11.1). If the raw data range is not between 1.0 and 999999.9, adjust so that all data falls within that range. If no data exist for the data point (e.g., different sample rates, no Move1 channel) write -9.9 as depicted below.

• Event Labels for Answers shall be “-----Yes”, “------No”, or “-----Ans”

Appendix D. Example Output File

Name of this file: D&-X20190812-1.01A
Source file: X20190812
Instrument: Lafayette Windows
Software Version: 11.8.3.218
Chart Date: 31Dec69
Time: 19:00
Examination Number: 1
Chart Number: 1
Number of questions: 9
Fastest Sample Rate (Hz): 30
Number of samples: 7907
Number of channels: 7
Sample Rate (Hz)- UPneumo: 30
Sample Rate (Hz)- LPneumo: 30
Sample Rate (Hz)- EDA2: 30
Sample Rate (Hz)- Cardio1: 30
Sample Rate (Hz)- Move1: 30
Sample Rate (Hz)- EDA2: 30
Sample Rate (Hz)- PL: 30

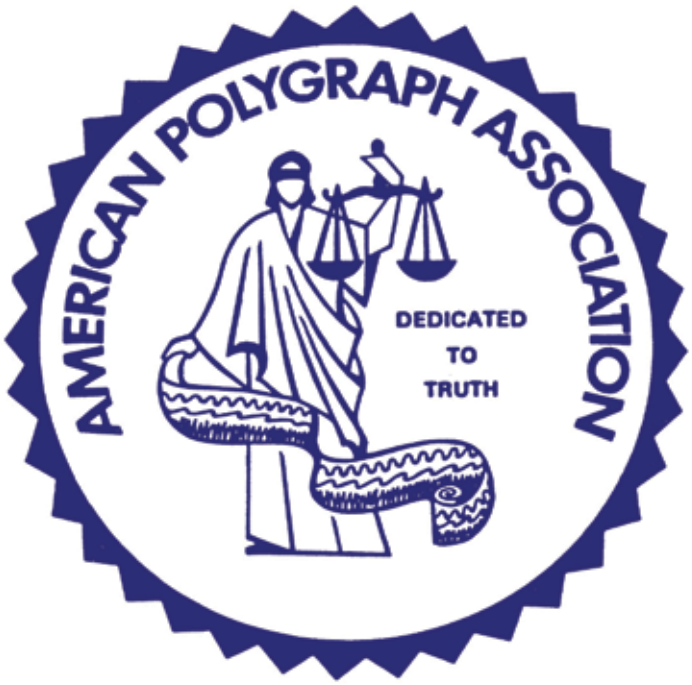
Event	Label	Statement
01	X	This practice test is about to begin. Please sit still. Look straight ahead. Listen carefully to each question and answer just as we have discussed. No other talking, and do not move during this practice test.
02	1	Did you write the number 1?
03	2	Did you write the number 2?
04	3	Did you write the number 3?
05	4K	Did you write the number 4?
06	5	Did you write the number 5
07	6	Did you write the number 6?
08	7	Did you write the number 7?
09	XX	This practice test is complete. Please sit still until I release the pressure in the cardio sensor. Name of this file: D&-X20190812-1.01A

Source file: X20190812

Event	Label	Statement
01	X	This practice test is about to begin. Please sit still. Look straight ahead. Listen carefully to each question and answer just as we have discussed. No other talking, and do not move during this practice test.
02	1	Did you write the number 1?
03	2	Did you write the number 2?
04	3	Did you write the number 3?
05	4K	Did you write the number 4?
06	5	Did you write the number 5?
07	6	Did you write the number 6?
08	7	Did you write the number 7?
09	XX	This practice test is complete. Please sit still until I release the pressure in the cardio sensor.

Event	Label	Begin	End	Answer
01	X	738	868	
02	1	1623	1678	1703
03	2	2386	2439	2466
04	3	3163	3222	3248
05	4K	3948	4004	4037
06	5	4735	4790	4813
07	6	5513	5568	5593
08	7	6303	6358	6383
09	XX	7081	7155	

Sample	Time	Label	UPneumo	LPneumo	EDA1	Cardio1	Move1	EDA2	PLE1
1	00:00.00	--------	16102.0	43230.0	29783.0	257758.0	463987.0	6794.0	237588.0
2	00:00.03	--------	16451.0	44021.0	29742.0	266777.0	463308.0	6793.0	600591.0
3	00:00.06	--------	17107.0	45612.0	29677.0	267911.0	462805.0	6787.0	871068.0
4	00:00.10	--------	18982.0	47430.0	29637.0	266684.0	463223.0	6777.0	898617.0
5	00:00.13	--------	20173.0	49071.0	29581.0	261089.0	464169.0	6764.0	805945.0
6	00:00.16	--------	21276.0	50545.0	29538.0	253123.0	464788.0	6748.0	709585.0
7	00:00.20	--------	22348.0	51505.0	29503.0	246371.0	464590.0	6729.0	620585.0
8	00:00.23	--------	23370.0	53670.0	29460.0	243706.0	464749.0	6709.0	533106.0
9	00:00.26	--------	24960.0	55452.0	29429.0	245422.0	464473.0	6688.0	477644.0
10	00:00.30	--------	26133.0	56988.0	29381.0	247821.0	464840.0	6665.0	471043.0
11	00:00.33	--------	27073.0	58639.0	29347.0	247632.0	465423.0	6640.0	482089.0
12	00:00.36	--------	28324.0	60031.0	29317.0	244604.0	465981.0	6614.0	470864.0
13	00:00.40	--------	29577.0	61346.0	29270.0	237597.0	466751.0	6588.0	424666.0
14	00:00.43	--------	30845.0	63759.0	29241.0	229131.0	467736.0	6561.0	359372.0
15	00:00.46	--------	32523.0	65201.0	29221.0	219407.0	468025.0	6534.0	292193.0
16	00:00.50	--------	33626.0	66480.0	29195.0	210471.0	468401.0	6507.0	238304.0
				DATA OMITTED HERE					
7903	04:23.39	--------	68276.0	48981.0	19955.0	109743.0	481490.0	14182.0	421374.0
7904	04:23.43	--------	66294.0	47338.0	19934.0	112466.0	480802.0	14089.0	409459.0
7905	04:23.46	--------	64283.0	44937.0	19919.0	112251.0	479897.0	14002.0	429179.0
7906	04:23.50	--------	63053.0	43162.0	19923.0	109281.0	479256.0	13919.0	436729.0
7907	04:23.53	--------	-9.9	41599.0	-9.9	-9.9	-9.9	-9.9	-9.9


AMERICAN POLYGRAPH ASSOCIATION
DEDICATED
TO
TRUTH