RECOMMENDATION TO ADOPT STANDARDIZED TEST DATA ANALYSIS SYSTEM

A Seminar Paper

Presented to

The Graduate Faculty

University of Wisconsin – Platteville

In Partial Fulfillment of the

Requirements for the Degree

Master of Science

In

Criminal Justice

By

Steven M. Shannon

2020

Under the Supervision of Dr. Susan Hilal, Professor UW-Platteville Criminal Justice Department

ABSTRACT

RECOMMENDATION TO ADOPT STANDARDIZED TEST DATA ANALYSIS SYSTEM

Purpose

Polygraph testing has been used extensively by local, state, and federal law enforcement agencies as well as by the intelligence community, yet no single scoring method has been recommended or adopted, to create an industry standard. By comparing the 7-Position system, 3-Position system, and the Empirical Scoring System (ESS) against one another, a formal recommendation can be made to the National Center for Credibility Assessment (NCCA) regarding adoption of a standardized Test Data Analysis (TDA) method. This recommendation may include the integration of the ESS into their curriculum. The review could also provide a basis to make a recommendation of a particular scoring system to become the national standard. If a recommendation is made to integrate and incorporate ESS into basic curriculum, it would then allow federal intelligence and law enforcement agencies to utilize ESS for TDA. If evidence supports the ESS as a more robust method, a recommendation that it should be adopted and used by the federal polygraph community will occur. A recommendation for further research that needs to be conducted will be included. The use of the most efficient scoring method will be recommended in order to help government agencies save time, money and resources.

Methods

The information used for this study will include an examination of secondary sources. These will include organization or agency websites, accredited journals, government reports, and the American Polygraph Association. It will include data from the federal polygraph program handbook published by the National Center for Credibility Assessment (NCCA). Studies that

review individual hand and computer scoring techniques will be analyzed and assessed against current scoring techniques being used. The scoring system(s) that have the highest level of accuracy and validity will be recommended for adoption on a nationwide basis when evaluating physiological data derived from any comparison question test (CQT) format.

Findings

Based off of previous empirical studies, no individual Test Data Analysis (TDA) system showed to be the most effective or robust for the hand-scoring of recorded physiological data. All three primary systems evaluated had very similar accuracy rates, inconclusive results, and rates of error. As a result, no particular system should be recommended for adoption as the national standard and additional future research should be pursued. Specifically, future research should include several large-scale studies that focus on criterion accuracy, effects of overall workload placed on examiners, and false negative error rates, especially as they relate to the Empirical Scoring System (ESS). The ESS should be given special focus since this paper supports the findings of numerous previous studies that the ESS shows promise as an effective system while providing users and consumers of polygraph data additional benefits that other systems cannot provide. The search for a TDA system that can be used as the national standard should continue until a recommendation can be made and adopted by the National Center for Credibility Assessment (NCCA).

ACKNOWLEDGEMENTS

First, I need to thank my wife and kids for their constant support and words of encouragement. You have enabled me to be successful in the polygraph profession as well as my graduate degree studies. I am eternally thankful for all of your support. All of you make the extra effort worthwhile.

Second, I need to thank all of the faculty and staff at UWP. The professors and the distance learning staff have all been extremely helpful and supportive. They have all had a profound impact on my academic career.

Third, I must thank Dr. Hilal for her encouraging words and assistance not only as my advisor for this seminar paper, but as a fantastic professor in other classes. The feedback from Dr. Hilal has been crucial in assisting me in not only completing online course work, but has helped me elevate my writing skills. She was able to help guide me through this seminar paper process eventually leading to a graduate degree, for that I am forever grateful.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION

Introduction

Currently, polygraph is used by federal, state, and local law enforcement as well as intelligence agencies across the nation. Collectively, thousands of polygraph exams are administered each year for pre-employment screening, witness or source testing, and criminal matters. There are currently 26 federal polygraph programs (divisions) that are contained within nine federal agencies (Robertson, 2012).

The results of polygraph exams can have profound effects on the examinees; including the loss of a government job or a potential prison sentence. Controversy over the use of polygraph is ongoing, but the usefulness of polygraph examinations is widely recognized (Nelson, 2016; Warner, 2005). The study of diagnostic features recorded by the polygraph has spanned nearly a century. Those features that are currently used for evaluation by the polygraph community have received empirical support (Nelson & Handler, 2010).

In 1988, the federal government restricted the use of polygraph exams for hiring in most private businesses through the Employee Polygraph Protection Act (Handler, 2015), leaving its use limited to a pre-employment screening tool to eliminate undesirable candidates, grant security clearances and enable access to specific classified programs. Yet, the federal government, nor any other governing body, has not required agencies to use any one standardized test data analysis (TDA) technique. As a result, there is no industry standard for hand-scoring polygraph data.

There is a variety of testing techniques and tools designed to conduct TDA that are recognized by the American Polygraph Association, yet that association has not endorsed or

standardized any specific protocol (Gougler et al., 2011). There are currently three primary systems of assessing the physiological responses generated during polygraph testing in federal government programs. They include: Global Analysis, 3-Position, and 7-Position systems. However, a fourth system, The Empirical Scoring System (ESS) is also touted as a highly effective and simple TDA method supported by empirical evidence (Nelson et al., 2011). Unlike the other systems, the ESS allows for the calculation of statistical significance of error rates for manual cutoff scores (Nelson & Handler, 2010). Improving inter-scorer reliability and establishing error estimates was a major goal in the development of the ESS (Krapohl, 2010).

Statement of the Problem

Currently, individual examiners and agencies do not have the ability to reference a single scoring method as being most effective. Literature challenging the accuracy of the 7-Position system, relative to other scoring methods (i.e. Backster's), have proven ineffective (Merion et al, 2008). Krapohl, Stern, and Brokema (2009) argues that the 7-Position polygraph scoring system is more accurate than other methodologies including the rank order approach. As a result, Department of Defense (DOD) uses this method of TDA (DOD, 2006a; DOD, 2006b) for its polygraphs and curriculum taught at the National Center for Credibility Assessment (NCCA). The ESS is widely used and taught worldwide; however, it is not taught or recognized for use by federal government entities.

A thorough review of the 7-Position, 3-Position and ESS has not been published by either a federal government agency or one of the major national polygraph associations. A thorough review of the four major scoring systems is needed to make a formal recommendation to NCCA regarding their TDA curriculum, including the possible inclusion of the ESS. Empirical studies (Krapohl, 2010; Nelson et al., 2011) indicates that the ESS is more efficient (requires less time to

conduct), more reliable (i.e., increased interrater agreement), reduces proportion of inconclusive decisions, and increases accuracy rates for truthful individuals, relative to the 7-Position scale. In a study produced by a federal law enforcement agency, 300 screening polygraph examinations were analyzed by applying the 7-Position and ESS approach. Results demonstrated a very high proportion of decision agreement, indicating ESS and 7-Position systems yield extremely similar decision outcomes (Comparative Analysis Using DHS Screening Data, 2018).

A thorough review of research available on global analysis, 7-Position, 3-Position, and the ESS must be completed in search for the most effective and efficient method of conducting test data analysis, which could lead to the recommendation of an industry standard.

Method of Approach

The information used for this study will include an examination of secondary sources. These will include organization or agency websites, accredited journals, government reports, and the American Polygraph Association. It will include data from the federal polygraph program handbook published by the National Center for Credibility Assessment (NCCA). Studies that review individual hand and computer scoring techniques will be analyzed and assessed against current scoring techniques being used. The scoring system(s) that have the highest level of accuracy and validity will be recommended for adoption on a nationwide basis when evaluating physiological data derived from any comparison question test (CQT) format.

Significance and Implications

In order for any federal, state, or local agency to be in a position to defend their polygraph program and the results of tests administered by those programs, they must rely on research and accreditation by related governing bodies. As law enforcement and intelligence agencies are scrutinized over the use of polygraph exams, they should have standardized

methods of conducting and scoring those exams. Hand-scoring physiology is subjective in nature and remains a potential liability issue for agencies. This becomes even more challenging in court, when agencies are asked to divulge p-value error rates in court (Nelson et al., 2011). The failure to standardize scoring methods could result in further limited use of polygraph.

This paper could be the first to compare all four primary scoring systems in order to recommend a "best practices" method for hand-scoring the physiological data collected during polygraph exams. By compiling research on the primary test data analysis systems, including the ESS, a formal, evidence-based argument regarding the most effective TDA system can be produced. This argument will result in a recommendation to NCCA regarding the inclusion of the most effective system, (if applicable) into their curriculum. By extension, this would allow US Government polygraph programs to use the most robust system of TDA available. The impact of this effort could be to save government agencies time and resources, while not sacrificing accuracy or endangering national security.

Assumptions

The first major assumption that must be made is that local, state, and federal polygraph examiners are taught and utilize similar physiological criteria for conducting TDA. Second, there will be an assumption made that those polygraph examiners, regardless of their training, use one of the four primary scoring methods illustrated above. These assumptions are necessary to establish a baseline and form a foundation for analysis. This analysis will guide a formal recommendation to polygraph professionals as we work to standardize processes within our profession.

Limitations

Some of the limitations within this paper are related to the assumptions. Many assumptions made above generalize polygraph training or schools that are attended by various polygraph examiners. Those examiners work in all 50 states, for both local, state, and federal law enforcement agencies as well as the private sector. Specifically, many of the recommendations made will need further research to determine their applicability in the private sector, especially in circumstances where loss of proprietary information may be the focus. The current research focuses primarily on federal regulated polygraph program instruction which has different demands than non-federal agencies. Finding ways to apply scoring techniques that are adopted or used by federal government agencies in non-federal agencies or the private sector will undoubtedly have challenges. Lastly, the most prominent limitation to this paper is that the research is dated and there are not many recent studies available on many of the topics.

CHAPTER II: REVIEW OF THE LITERATURE

This literature review is comprised of five sections. First, a brief review of the evolution of polygraph is provided, including its use in criminal and pre-employment screening situations. Second, is a discussion of the three primary physiological channels in which data is collected in all polygraph exams, as well as diagnostic features contained in those channels that have empirical support. Third, is a discussion of the most common scoring systems used in the US during polygraph testing. Fourth, the common arguments against the use of polygraph are reviewed. Lastly, an explanation for the differences between different federal agencies polygraph programs is provided.

History

Dr. William Marston first devised the use of physiology to detect deception in 1915, during World War I. At that time, he was tasked with creating a technique in which to question prisoners of war. His first real-world case was an attempt to identify a suspect in the theft of a codebook from the U.S. Surgeon General's office. Marston utilized a standard blood pressure cuff to measure the systolic blood pressure during questioning. He believed that by monitoring the systolic blood pressure, he would be able to identify deception (Alder, 2007).

In 1921, John Larson, a policeman and physiologist working in Berkeley, CA would use Marston's findings to create an apparatus that simultaneously measured changes in blood pressure, heart rate, and respiration rate in order to detect deception (Larson, Haney, & Keeler, 1932). An Italian psychologist named Vittorio Benussio had also previously published his finding on the respiratory symptoms of deception which aided Larson in the development of his apparatus (Alder, 2002).

Most of the early research conducted on polygraph was accomplished by Larson during the 1920s while he was a police officer for the Berkeley Police Department. Larson's Chief of Police, August Vollmer, believed Larson's work could significantly improve the effectiveness of his department in their law enforcement investigations. As a result, Chief Vollmer allowed Larson to test and refine his polygraph by using it in real criminal cases. Vollmer's focus was almost solely on the practical value of the polygraph, something that most law enforcement would end up supporting (Synnott, Dietzel, & Ioannou, 2015).

Larson also benefited from the help of his aid and protégée, Leonard Keeler. Keeler is often credited with creating the first polygraph testing procedures which included the Relevant/Irrelevant Question Technique (Keeler, 1930). Keeler is also the individual who made the polygraph a portable apparatus and possibly more importantly, added the galvanic skin response (GSR) or EDA channel to it in 1938, which today is touted as the most diagnostic physiological channel used in polygraphs. Keeler added this diagnostic channel as a result of work conducted by psychologist Reverent Walter G. Summers. At that time, Summers was working at the Fordham University Graduate School (Summers, 1936). EDA as well as the other physiological channels will be thoroughly reviewed in subsequent sections.

Keeler was the first to patent his polygraph as his focus was financial and commercial rather than academic. Keeler was also the first to establish a polygraph school (Alder, 2007). After Keeler's death in 1949, the history of polygraph continued unabated with John E. Reid. Reid is possibly best known for his controversial Reid Technique of interview and interrogation (Gudjonsson, 2003). Reid also established his own polygraph school as well as the Comparison Question Test. This testing format replaced Keeler's Relevant/Irrelevant technique as the most popular technique, which it remains today (Raskin & Honts, 1987).

Neither Keeler nor Reid were concerned with a standardized approach to polygraph or the peer review of polygraph research. Rather, they invited serious criticisms of polygraph based on their claim that polygraph was an interrogative device designed to assist examiners in a clinical interpretation of the examinees behavior, and the "real lie detector" was the examiner using their training and experience (Reid & Inbau, 1977). Since this approach attracted serious criticisms of polygraph, Cleve Backster, a student of both Reid and Keeler realized a numerical scoring system would be beneficial. As a result, Cleve Backster's numerical scoring system was the first of its kind to be used for numerically evaluating polygraph charts. His system solely relied on information from charts making the evaluation more scientific and objective (Grubin & Madsen, 2005).

Diagnostic Physiological Criteria

Electrodermal Activity

Electrodermal Activity (EDA), sometimes called galvanic skin response (GSR) is the physiological channel added to polygraph by Keeler in 1938 (Synnott, Dietzel & Ioannou, 2015). Electrodermal activity, in its simplest form, describes the electrical conductance of skin in response to sweat secretion by eccrine sweat glands. When a low constant voltage is applied, the change in skin conductance can be measured non-invasively (Fowles et al., 1981). This allows a wide application of skin conductance measures to be used in both basic and clinical research (Benedek & Kaernbach, 2010).

The changes in electrical properties of the skin are caused by the electrolytes contained in sweat that is secreted by the eccrine sweat glands. Eccrine sweat glands are exclusively innervated by sympathetic activity from the autonomic nervous system (ANS), a common theme amongst most of the physiological channels collected during a polygraph. The ANS controls the

body's auto regulatory processes including temperature, heartrate, and blood pressure to ensure homeostasis is maintained. The ANS is essentially broken up into two parts, sympathetic activity and parasympathetic activity. Sympathetic activity includes activity that is created by the fight-or-flight response. This activity is indicative of bodily arousal resulting from emotional expressions (Topoglu et al., 2020).

Electrodermal activity is important since it is a concomitant of the orienting response as well as general emotional arousal. EDA however is not indicative of any specific type of emotional arousal. Both positive, happy thoughts and negative or threatening stimuli can result in additional arousal, thereby increasing the skins conductance. This means that GSR does not represent any particular type of emotion, rather it is only a measure of the intensity of whatever type of emotion the individual is experiencing (Critchley, 2002).

Eccrine sweat glands are activated by the postganglionic sudomotor fibers. These fibers receive signals through the ANS and are responsible for innervating the sweat glands. Each sudomotor nerve burst corresponds to an observable skin conductance response. The amplitude of the nerve burst is linearly related to the number of sweat glands that have been activated. As a result the skin conductance response amplitude (height) is considered a measure of sympathetic activity (Freedman et al., 1994).

EDA data acquisition is accomplished during polygraph testing by attaching electrodes to areas of the skin which have high concentrations of eccrine sweat glands. The electrodes that are attached passively monitor the electrical properties between the skin and the components that are attached to the examinee. The location where these electrodes can be attached to an examinee varies but the most accurate EDA data is measured at palmar sites on the hand. These areas include the medial and distal phalanges of the middle, index, and ring fingers as well as areas of

the palm. The wrist and feet can also be used to collect EDA since these regions have higher eccrine sweat gland densities compared to the remainder of the body (Topoglu et al., 2020). The below figure depicts some of the typical measurement sites for EDA.

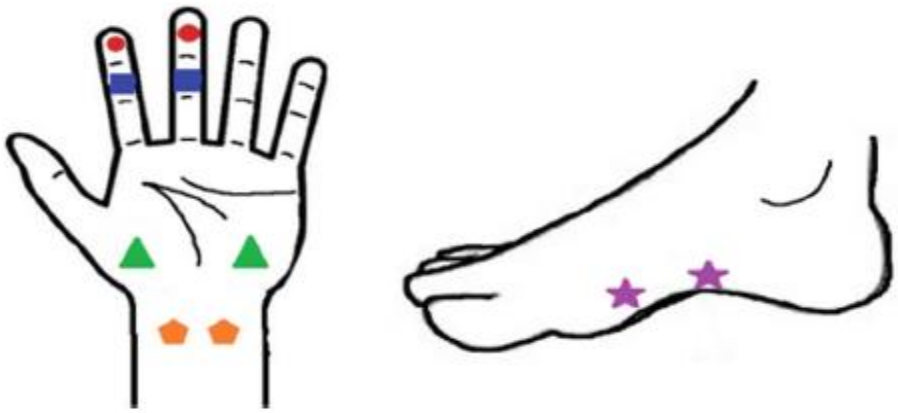


Fig. 1. Typical measurement sites for EDA.

(Topoglu et al., 2020)

There are two types of activity that occurs within EDA. Phasic activity and tonic activity. Phasic activity is frequently referred to as the skin conductance response which shows fast fluctuations caused by sympathetic arousal resulting from some type of stimuli (Topoglu et al., 2020). A skin conductance response that is related to some specific event, like a question posed during a polygraph exam, allows the examiner to capture the resulting stimulus response. A typical skin conductance response will have some latency between the actual application of stimuli and the peak of the response. The delay in amplitude increase when recording phasic activity within EDA is from stimulus onset to response onset, typically one to three seconds in length (Topoglu et al., 2020).

Cardiovascular

Cardiovascular arousal is one of three primary channels of physiological data that is recorded within a polygraph test. The cardiovascular channel describes the monitoring and recording of heart rate and blood volume changes. Typically, the brachial artery is the primary source of physiological data. The brachial artery extends down the arm and eventually branches near the elbow into the ulnar and radial arteries. The brachial artery is a reliable source of cardiovascular activity since it is one of the largest blood vessels within the human body (Moore, 2017). The brachial artery is also the most frequently accessed artery to gain blood pressure information by the medical community (Krapohl & Shaw, 2015).

A cardio vascular monitor (CAM), originally created for the Air Force was designed to monitor the cardiovascular activity of an individual for a long period of time. The same chart tracings that used to be produced by a CAM are now produced by a conventional cuff (Davidson, 1979).

Today, cardiovascular activity is commonly monitored by placing a conventional blood pressure cuff around the arm where the brachial artery is located. The air bladder contained in the cuff is then inflated using a hand pump. Pressure within the air bladder pushes the artery against the bone in the arm and then once pressure is set appropriately, polygraph instrumentation can pick up changes within the cardiac cycles. Blood pressure dynamics associated with polygraph can then be recorded (Krapohl & Shaw, 2015).

Respiration

Respiration within polygraph describes recording the movement of the chest and abdomen that is caused as a result of ventilation within the human body. This movement is

typically captured in polygraph by affixing rubber tubes to the examinee that are capable of expanding and contracting during respiration. These rubber tubes are commonly placed around the chest of the participant with one just above the subject's heart, and the other just below the subject's heart (Krapohl & Shaw, 2015). These components record the rate of respiration and depth of respiration during testing (Slavkovic, 2018)

The pneumograph channel is one of the most difficult channels to evaluate for a polygraph examiner. There are several factors that can alter this channel making it more complex. First, the pneumograph channel is the only channel that is not completely controlled by the autonomic nervous system allowing an individual to consciously control their breathing. Secondly, it is affected by numerous naturally occurring human behaviors. These include sniffing, coughing, sneezing and swallowing among others. Lastly, just the act of verbally answering the questions during a polygraph examination can alter the data (Krapohl & Shaw, 2015).

Empirically Supported Diagnostic Features

Certified federal polygraph examiners are trained to evaluate or use 20 types of changes in polygraph data to make a determination of an individuals' deception status. Computerized scoring programs are capable of analyzing as many as 85 variations in polygraph data (Kristjansson et al., 2006). However, not all of those diagnostic features have been supported by empirical evidence. The following will review and summarize those features that have empirical support for their diagnostic value, and as a result, are the recommended and adopted features taught to federal polygraph examiners.

Respiration

The measure of respiration line length (RLL) encompasses several diagnostic features within the channel. Those features with empirical support include apnea, suppression, progressive decrease in amplitude, decrease in rate, and change in the inhalation/exhalation ratio.

There are several different types of polygraph exams that can be administered depending on circumstances as well as agency policy. Three major types of polygraphs conducted include the directed lie and probable lie comparison question tests and the relevant/irrelevant test. Horowitz et al. (1997) conducted a study to determine the effectiveness of probable lie tests when compared to trivial directed lie, personal directed lie, and relevant/irrelevant polygraph exams.

The study showed that relevant/irrelevant tests provided an unacceptably high rate of false positive test results. That comparison questions have different psychological and psychophysiological meaning for individuals and are important in the physiological detection of deception. Numerical scores obtained were similar between the directed lie and probable lie exams. Respiration data was found to be strongly in the predicted direction, but only when probable lie participants were being tested. As a result, the study found that respiration might be the least reliable diagnostic channel when numerically scored. Respiration line length also had the largest drop in validity when cross-validated by the computer scoring model (Horowitz et al., 1997). More recently, a study found that respiration line length calculations are biased due to respiration cycle characteristics during time intervals. By using a weighted average method, performance based on respiration line length increased during Concealed Information Tests (CIT). This weighted average method is promising in assessing respiration changes more accurately (Matsuda & Ogawa, 2011).

In a study conducted in 2016, RLL was shown to be effective at identifying concealed information during a Concealed Information Test (CIT). A CIT is a method used in psychophysiological detection that is often useful in criminal investigations. A typical use is to help detect whether a person is concealing information they wish not to disclose. For example, it can be assumed that only guilty persons will know the details of particular crimes so enhanced physiological responses from critical stimuli indicates recognition, thus indicating possible guilt. An innocent suspect will see all stimuli as neutral and respond in a non-consistent manner (Elaad, 2016).

Participants in the CIT study were given course credits and a possible monetary bonus for participating and successfully preventing disclosure of their concealed information. During subsequent polygraph exams, skin conductance, respiration, and cardiovascular activity were monitored and recorded. Respiration responses were defined by total RLL during a 15 second interval following the onset of stimuli. The responses to all items were assessed and the largest response was assumed to be the detected item (Elaad, 2016).

When skin conductance was used alone, three of the six items were correctly identified. RLL proved to be more effective which allowed four of the six items to be correctly identified. When RLL was combined with the skin conductance response measured by amplitude, five of the six critical items were correctly identified (Elaad, 2016). These findings underscore years of research that shows the effectiveness of RLL in detecting deception.

Cardiovascular

There are several cardiovascular features that have empirical support for their ability to help detect deception in polygraph. Those features include cardiovascular arousal, cardiovascular

duration, and cardiovascular pulse rate changes. These features are typically monitored or recorded by using a blood pressure cuff however, other components such as a photoplethysmograph or electrocardiogram leads can also be used (Kircher et al., 2010).

In one study, 417 community members were recruited to participate in a mock crime experiment. Those individuals were eventually given a polygraph exam in an effort to detect deception amongst the participants. Each participant was paid $30 to participate with the opportunity to earn an additional $50. The study had three main objectives. First, it focused on the effectiveness of a stimulation test and feedback prior to the actual examination. Second, it reviewed the relative effectiveness of a probable-lie and directed-lie comparison question test. Lastly, it assessed the ability of new physiological measures to detect deception while collecting standard physiological data (Kircher et al., 2010).

Kircher et al. (2010) used a cardiograph to measure cardiovascular arousal during polygraphs related to the mock crime. The strength of the cardiovascular response during question presentation was measured by amplitude and duration. Results found increases in the baseline of the cardiograph tracing were diagnostic when evaluating an individual for deception.

Cardiovascular duration has also been investigated by many to determine its diagnostic value. Honts, Raskin, and Kircher (1994) conducted a study to determine the effects of countermeasures on control-question polygraph tests. This study consisted of placing a help wanted add in two local newspapers and screening respondents initially by telephone. Ultimately, 120 subjects completed the experiment which entailed potentially being involved in a mock crime, receiving training on countermeasures, and the examination by polygraph to detect deception and the use of countermeasures. During the polygraph tests, 11 physiological changes were collected for evaluation including an increase in cardiovascular response amplitude and

duration. Those exams were scored by the original examiner and an independent evaluator. Computerized measurements were also obtained and analyzed using the same metric.

There were several important findings to note in this study. First, that interrater reliability was high, indicating that the use of numerical scoring is extremely reliable. This was accomplished by correlating the total numerical scores assigned by evaluators. Secondly, cardiovascular duration, the focus of this section, was found to be diagnostic when detecting deception (Honts, Raskin, & Kircher, 1994).

A study which compared human versus computerized lie detection also recruited subjects from a local community via a classified newspaper advertisement. In this study 100 subjects were randomly assigned a guilty or innocent condition and those assigned a guilty condition received tape recorded instructions and a set amount of time to commit a mock theft. All participants were given a polygraph exam by an examiner who did not know whether the examinee was guilty or innocent of the theft. The results were scored by experienced examiners as well as by computer (Kircher & Raskin, 1988).

The primary focus of the study revolved around human versus computerized scoring of polygraph data however the utility of different types of data should not be ignored. Ultimately the study found that definite decisions made by the original examiners were 95% accurate. The computerized scoring was slightly higher at 98% accurate. The results of the relative utility of physiological components showed that an increase in baseline and duration within the cardiovascular recording were useful diagnostic features when discerning truth from deception (Kircher & Raskin, 1988).

Electrodermal Activity

Three primary features of electrodermal activity include amplitude, duration, and complexity. Significant differences in amplitude and duration have been observed when skin conductance was compared between innocent and guilty control subjects. Countermeasures also appeared to have little affect when EDA responses were compared between guilty control and countermeasure participants (Honts, Raskin, & Kircher, 1994).

Kircher et al. (2010) tested if the pretest portion of a polygraph exam would improve the accuracy in either probable-lie or directed-lie tests. It also examined the effect of providing feedback to participants regarding the outcome of their test. The participants were paid to participate in a mock crime scenario with two levels of guilt and four variants of pretest procedures. Overall, half of the participants were innocent of the mock theft and the remainder were guilty. No significant differences were found between the two types of tests administered however important information about physiological features were realized.

Among those findings included that skin conductance was the most diagnostic measure when detecting deception. It also noted that the amplitude when measuring skin conductance is not only the typical measure used, but it is also the most valid measure. As a result, the skin conductance and skin potential findings were especially encouraging since different physiological mechanisms may be responsible for each type of responses (Kircher et al., 2010)

When the examination results (guilty or innocent) were reviewed and the different physiological recordings were analyzed, the skin conductance response was found to have the largest impact on decisions about guilt or innocence. Skin potential was found to be highly diagnostic of truth or deception in these mock crime scenarios. Skin potential was also highly

correlated with the measurement of amplitude in skin conductance. Ultimately, of all the measurements that are obtained during field polygraph examinations, the measurements of electrodermal activity carry the most weight in decision processes not only for expert polygraph examiners, but also computer models (Kircher et al., 2010; Kircher & Raskin, 1981; 1988; Raskin et al., 1988;).

Common Scoring Systems

The following will review the four most popular scoring systems used within the polygraph community. The first system, global analysis does not rely on the assigning of numerical scores. As a result, it will be briefly reviewed since it is primarily applicable to locating artifacts and one type of polygraph testing. It could not be recommended as an industry standard as it is not applicable to most criminal and pre-employment polygraph exams that utilize comparison question test formats. Therefore, any recommendation for adoption of a particular scoring system would be made with the intent to be used in evaluating polygraph data that has been derived from comparison question test formats. The potential scoring systems that could be recommended that rely on the application of numerical scores will be thoroughly reviewed. Scoring rules for the three most prominent systems as well as the potential numerical scores that can be assigned for each system will be reviewed in the ESS section.

The possible outcomes of a polygraph examination conduced in a specific-issue or criminal related manner are “NDI: no deception indicated” (passed), “DI: deception indicated” (failed) or “inconclusive.” Inconclusive tests occur when no opinion can be rendered regarding the outcome of the test, typically because a cutoff score was not achieved. In screening tests, NSR or no significant response signifies a pass, SR or significant response signifies a fail, and

inconclusive indicates a conclusive decision could not be reached (Department of Defense, 2006a).

Global Analysis

Global analysis is the primary method used for evaluating charts obtained during Relevant/Irrelevant (R/I) polygraph tests. The R/I screening technique is widely used, especially federal intelligence agencies and it does not have a validated scoring system (Krapohl & Rosales, 2014). R/I does not use comparison questions, which would allow for a comparison to be made to relevant questions and evaluate those reactions compared to one another. All numerical scoring systems including the 7-Position system are inappropriate for use in evaluating R/I, although it has been attempted (Krapohl & Shaw, 2015).

Examiners that administer R/I tests have created a mnemonic to refer to the way they evaluate physiological responses. Conspecnificance refers to consistency, specificity, and significance. Examiners look for reactions that consistently occur, are significantly greater than others and appear during a specific question (Krapohl & Shaw, 2015). Although global analysis is primarily used during R/I testing, it also serves an important purpose in evaluating comparison question test data.

Global analysis can assist examiners in effective chart interpretation. First, it allows the examiner to assess the tonic activity and how labile or dynamic any of the tracings may be. Secondly, examiners can evaluate the physiological data and exclude any that should be considered an artifact. Lastly, it allows the examiner an opportunity to evaluate whether the data has been manipulated by the examinee and how much weight to place on the data (Krapohl & Shaw, 2015).

7-Position System

Initial numerical scoring systems were first introduced by Cleve Backster and were a welcomed improvement over global approaches (Backster, 1969). These initial systems were very complicated as they contained numerous complex rules which had not been validated by research. Complex scoring systems are also likely to create unreliable outcomes since charts will likely be evaluated differently by different examiners (Kristjansson et al., 2006).

Possibly the best known system created by Backer is the 7-Position scale. This system is still widely found in the field of polygraph. Like the name implies this system has seven scores which an examiner can assign when evaluating relative activity observed at relevant questions, and by comparing that activity to the activity observed at adjacent comparison questions. The values of -3, -2, -1, 0, +1, +2, and +3 can be used. Negative values are to be assigned when the reaction to a relevant question is greater than to the corresponding comparison question. Conversely, a positive value is assigned when a reaction to a benchmark comparison question than that of the relevant question (Krapohl and Shaw, 2015).

Although different polygraph schools teach different rules, the recommended rule is to use the stronger of either adjacent comparison question, while scoring each channel independently. This means that a cardiovascular response can be scored against a stronger preceding comparison question while the EDA channel can be compared to a stronger comparison question that follows. Any comparison question test format that does not utilize the placement of a comparison question immediately before a relevant question creates a structural problem which affects decision accuracy (Krapohl & Shaw, 2015).

3-Position System

The 3-Position system like the 7-Position system is used to evaluate polygraph data obtained during comparison question formats. The responses at a relevant question are assigned a plus (+) value when a response at a comparison question is bigger. A minus (-) is assigned when the response at the relevant question is larger than a response at a comparison question. A zero is assigned when the responses at relevant and comparison questions have no apparent difference in magnitude (Department of Defense, 2006a).

There are only three numerical scores that can be assigned when using the 3-Position scale, therefore it is typically considered an abbreviated form of the 7-Position scale. These scores are -1, 0, and +1. All scores assigned using this system is based on the "bigger is better" principle meaning the response with the largest magnitude will be assigned the score (Krapohl & Shaw, 2015). For single-issue testing, the 3-Position scale is just as accurate at the 7-postiion scale when decision rules or cut-off scores are adjusted (Harwell, 2000).

Empirical Scoring System

Nelson et al., (2008) was the first paper to reference the empirical scoring system. This paper was followed shortly after by many others that were examining its accuracy and reliability for potential use in scoring different polygraph testing formats. As a polygraph scoring method, it is very similar to the 3-Position scoring system, yet it is among one of the most studied techniques (Krapohl & Shaw, 2015).

The ESS was designed to provide a validated and reliable scoring model that was based on the simplest solutions which have empirical evidence to support all of the assumptions, principles, and procedures utilized. The ESS conforms to valid principles and can be used with a variety of comparison question test formats including properly constructed single-issue and

multiple-issue screening examinations. ESS allows the user to select optimal cutoff scores based on normative data and the operational needs for resolution versus precision. By basing cutoff scores on normative data, the user can calculate the probability of an error in the test result (Nelson et al., 2011).

As reviewed earlier, scoring of polygraph data is the assignment of a numerical value based on a comparison of two separate responses, one at a relevant question and one at an adjacent comparison question. The primary different between ESS and the closely related 3-Position scale occurs in the assignment of a numerical value for the EDA channel as well as ESS decision rules, which will be reviewed later. Scoring EDA in the 3-Position scale only allows for the assignment of a numerical value of a -1, 0 or +1. ESS differs in this area as it puts more weight on the EDA channel. As referenced earlier, EDA has been found to be the most diagnostic channel for detecting deception. Therefore, one of the primary focuses of ESS to give twice the diagnostic value to the EDA channel as compared to the pneumograph channel or the cardiovascular channel. As a result, the only numerical values that can be assigned to the EDA channel when using the ESS is a -2, 0, or +2 (Krapohl & Shaw, 2015). This weighting scheme results in an increase in test sensitivity to deception, a reduction in the number of inconclusive test results, and no change in test specificity to truthfulness (Nelson et al., 2011)

Proper scoring of polygraph data entails the adherence to clearly defined rules. By following those rules, examiners are able to execute scoring in a reliable way which will produce consistent results that agree with the evaluation by a second examiner using the same method. Some of the validated rules include using a specific time period for evaluation referred to as the response onset window (ROW). Since examinees physiological responses will typically have some latency, there must be a defined period of time where a response can be evaluated by the

examiner. That period allows the examiner to have confidence that a specific response was elicited by a test question (Krapohl & Shaw, 2015).

The scores assigned primarily depends on the channel being evaluated, the scoring system being utilized, and the amplitude of the responses. The 7-Position scale relies on ratios when evaluating the EDA channel and ratings of responses for the cardiovascular channel. Potential scoring assignments for each channel are provided below and separated by the three numerical scorings systems under review (Krapohl & Shaw, 2015).

Table 1

Pneumograph Scores for 7-Position, 3-Position, and ESS

Ratio	Score		
	7-Position	**3-Position**	**ESS**
Noticeable	± 1	± 1	± 1
Significant	± 2	± 1	± 1
Dramatic	± 3	± 1	± 1

(Krapohl & Shaw, 2015)

Table 2

Ratios Used for EDA Score Assignment

Ratio	Score		
	7-Position	**3-Position**	**ESS**
Bigger is Better	± 1	± 1	± 2
2:1	± 1	± 1	± 2
3:1	± 2	± 1	± 2
4:1	± 3	± 1	± 2

(Krapohl & Shaw, 2015)

Table 3

Cardiovascular Scores for 7-Position, 3-Position, and ESS

Response	Score		
	7-Position	**3-Position**	**ESS**
Noticeable	± 1	± 1	± 1
Significant	± 2	± 1	± 1
Dramatic	± 3	± 1	± 1

(Krapohl & Shaw, 2015)

ESS includes different decision rules and procedural steps for categorizing scores and ultimately test results. These different rules and scores can be tailored to achieve changing operational objectives including sensitivity and overall accuracy. There are three primary scoring rules contained in ESS. First, the grand total rule involves the adding of all scores into a single

numerical result. This rule provides the simplest decision making and highest level of accuracy. The cost of this approach is a slightly higher inconclusive rate (Nelson et al., 2011).

The second approach is the spot score rule. This rule is used to assess results from a multiple-issue test like a screening exam. Using this rule involves calculating the subtotal scores for each relevant question. This approach can be used when a grand total approach has not lead to a conclusive result (Nelson et al., 2011).

Two-stage rules include the combined sequential use of the grant total rule and the spot total rule. If grand total scores are statistically significant to indicate a DI or an NDI test result, then that interpretation is correct. If the grand total results in a no opinion or inconclusive test result, the spot score rule is then applied. The use of the spot score rule, following the grand total rule does not allow for a solution to achieve NDI (Nelson et al., 2011).

Arguments against Polygraph

One frequently cited argument against the use of polygraph testing comes from the US National Research Council's report on the use of polygraph in personnel screening. This report was authored by the US National Research Council at the request of the Department of Energy and is the most extensive review of scientific evidence on polygraph (Synnott, Dietzel & Ioannou, 2015). In the report, the authors provided the following frequently cited conclusion:

> Notwithstanding the limitations of the quality of the empirical research and the limited ability to generalize to real world settings, we conclude that in populations of examinees such as those represented in the polygraph research literature, untrained in countermeasures, specific incident polygraph tests can discriminate lying from truth telling at rates well above chance, though well below perfection.

> Because the studies of acceptable quality all focus on specific incidents, generalization from them to uses for screening is not justified. (National Research Council, 2003, p. 4)

An additional argument against using polygraphs for screening tests, especially when CQT formats are utilized is that there are flaws contained in the underlying assumptions about how examinees will respond to different types of stimuli. Additional issues with the use of polygraph include that polygraph research is lacking in validity and scientific rigor (Synnott, Dietzel, & Ioannou, 2015).

Variable Program Factors

The National Center for Credibility Assessment (NCCA) is the governing body for polygraphs within the federal community. As a result, NCCA sets the standards on numerous aspects of polygraph testing to ensure polygraph is used in the most professional manner while ensuring it remains an effective investigative aid (Synnott, Dietzel, & Ioannou, 2015).

Although NCCA provides a handbook that is meant to guide the administration of federal PDD programs, this publication clearly allows agency heads to establish procedures regarding their own agencies PDD program. This includes which types of tests should be administered and which systems will be used in evaluating that data. As a result, some agencies can administer a certain type of polygraph test and if the applicant is unsuccessful in clearing the examination, they can be brought back for additional testing in which a different type of test and evaluation system could be used (Department of Defense, 2006b).

NCCA's handbook does require that each federal agency with PDD capability create a quality control process to ensure a review is completed according to federal regulations. However, agencies are allowed to decide if they should utilize polygraph based on their mission set, what their process should be for approving examinations if they have a PDD program, and how they review grievances that are filed based on their PDD program (Department of Defense, 2006b).

The primary factor affecting whether a federal agency utilizes a polygraph program comes from an undated memorandum signed by President Johnson. Under the terms of that memorandum, agencies within the executive branch could not use polygraph examinations for screening, personnel investigation, or for intelligence or counter-intelligence operations. The exception to this rule which was included in the memorandum states that the Chairman of the Civil Service Commission (now the Office of Personnel Management (OPM)) must have certified the agency as having an intelligence or counterintelligence mission that directly affects national security. The agencies that received this certification then had to prepare regulations that governed the use of polygraphs. Those regulations among other things, had to specify the specific purpose for the use of polygraph, a directive that informed the person that a polygraph would be utilized with as much advance notice as possible, that voluntary consent would be received prior to examination, and that the questions to be asked had to have specific relevance to the subject of the inquiry ("Use of Polygraph Examinations in the Department of Justice", 2006).

Conclusion

The three primary scoring systems used today in evaluating physiological data have different decision rules, cutoff scores, and levels of accuracy. Some differences include the ability to assign different numerical values based on ratios or levels of response observed within physiological channels. Simplicity in scoring rules is important because it decreases disagreements between examiners and are allows them to be more easily explained to non-polygraph professionals (Nelson et al., 2011). Even though some scrutinize the continued use of polygraph and the research that forms its foundation, polygraph research has continued unabated and the use of polygraph examinations is expected to increase in the future (Synnott, Dietzel, & Ioannou, 2015).

CHAPTER III: ANALYSES OF SCORING SYSTEMS

In order to determine which scoring system would best fit the needs of most users and consumers of polygraph data, a thorough comparison of the three primary systems is required to review strengths, weaknesses and accuracy levels of each respective system and is provided below.

Global Analysis

Global Analysis, when used by interpreters has produced accuracy levels (M = 87.4%) which are significantly lower than the levels of accuracy achieved by the same interpreters using the seven-position system (M = 98.9%). The most significant difference between the global system and the use of numerical systems was with regard to innocent suspects. In one study, global evaluation was responsible for a high percentage of false positive errors (26.4%). This number was over seven times the percentage of false positive errors for all numerical evaluations which was 3.6%. This study alone suggests that restricted, systematic approaches to chart evaluation is more accurate than subjective global evaluations (Kristjansson et al., 2006).

7-Position Scoring system

This scoring system has generally produced good overall levels of scoring accuracy, especially when used on Zone Comparison Tests (ZCTs) and Modified General Question Test (MGQT) examinations. Sometimes referred to as the Utah Numerical Scoring System, the simplified 7-Position scale (compared to Backer's) has produced an accuracy level of 82.3% when evaluating ZCT and MGQT examinations. The same data showed statistically significant results when assessing the overall interrater agreement. The rate of concurrence amongst examiners were highest using the 7-Position scale and it ranges from 78.8% to 92.0% for the

MGQT examinations. The range of concurrence for ZCT exams was slightly lower at 76.8% to 81.0% (Blackwell, 1998).

When examiners used the 7-Position scale, they yielded correct decisions in 66.8% of the cases, were incorrect 4.2% of the time, and had a no opinion rate of 29.0%. When no opinion exams were removed, the raters using the 7-Position scale averaged 94.1% correct decision outcomes (Blackwell, 1998).

Similarly, an additional study found the 7-Position scale yielded correct decisions 68.5% of the time when using cutting scores of +/-6. Decision outcomes were incorrect 2.6% of the time and 28.7% of the cases resulted in no opinion outcomes. Using 7-Position scales, proportion of correct decisions was much greater than chance. When no opinion results were excluded, scorers averaged 96.3% correct decisions (Harwell, 2000).

In numerous studies (Honts et al., 1994; Horowitz et al., 1997; Kircher & Raskin, 1988; Podlesny & Raskin, 1978; Rovner et al., 1979), the 7-Position system (Utah Scoring System) showed a high level of agreement on decisions between the original examiner and an independent evaluator. Those levels ranged from 96% to 100% agreement and the percentage of agreements exceeded 95% when both the original examiner and the independent evaluator both had come to a definitive decision. The correlation between the numerical scores of the independent and original examiner ranged from .92 to .97. The validity of this system has also been demonstrated by many of the same studies with the addition of Raskin and Hare (1978). The combined findings from all of these studies showed a high level of decision accuracy in laboratory settings. When inconclusive outcomes were excluded, the overall percentage of correct decisions on guilty subjects was 91%. The overall percentage of correctly identified innocent subjects was 89% (Bell et al., 1999).

3-Position Scoring System

Commonly referred to as a simplified version of the 7-Position system, the 3-Position system eliminates the need for an examiner to determine or estimate how much bigger a response is than the response at an adjacent question. Rather, the "bigger is better" rule is applied by eliminating complexity of rendering decisions. This is one advantage that the 3-Position system has over the 7-positon system. This simplification helps reduce variability amongst evaluators' and was thought to increase interrater scoring reliability which is an important factor when evaluating the effectiveness and efficiency of a scoring system. A significant increase in the interrater reliability was not observed when the 3-Position system was compared to the 7-Position scale. This is not a surprising result since 90.7% of the scores assigned using the 7-Position scale ranged from -1 to +1, the same scoring range the 3-Position system uses (Krapohl, 1998).

Krapohl (1998) found that the 3-Position system has some advantages over the 7-Postion scale. Specifically, the 3-Position system correctly identified 14 more innocent programmed examinees and resulted in 17 less no opinion decisions out of 250 decisions than that of 7-Position scale. This resulted in a six percent (6%) reduction in the number of inconclusive tests and a 6.6% increase in the number of correct innocent identifications made by the examiners. Most of the information about the 3-Position scale comes from studies which included a comparison to the 7-Position system. Krapohl (1998) points out that over 90% of scores assigned using the 7-Position system are the same numerical values that would be assigned if the examiners used the 3-Position system. This overlap seems to contribute the lack of available research that specifically studies the 3-Position system independently. Based on the literature

available, the findings of such studies would be very similar to those that have been done on the 7-Position scale.

Empirical Scoring System (ESS)

The empirical scoring system (ESS) is an example of how an evidence-based approach fosters simplification and hones the field practices of polygraph to the necessary robust essentials. One example of that simplification in polygraph scoring was a reduction in the number of physiological reaction features that could be scored. This reduction eliminated features that lacked scientific support and ultimately reduced the number from 23 to 12 (Department of Defense, 2006a; Department of Defense, 2006b). The ESS further reduced the number of features evaluated from 12 to 3 (Nelson, Handler, & Senter, 2012).

Several studies have shown that the ESS achieves high levels of criterion accuracy. The mean criterion accuracy is usually above 90% with inconclusive rates being between 10% and 20% (Nelson et al., 2011; Nelson et al., 2016; Nelson, Handler & Senter, 2012). The ESS has also demonstrated high levels of decision agreement amongst examiners when inconclusive tests are removed. The level of decision agreement even amongst inexperienced examiners has demonstrated an average rate of agreement of 85% (Handler et al, 2010). Accuracy rates and interrater reliability are only two of the strengths of the ESS. Additional benefits of the ESS include accelerated skill acquisition, increased skill retention, and increased generalizability of experimental results to field settings (Nelson et al., 2011). One of the most promising aspects of the ESS is the ability to assign associated p-values and normative data in regards to specificity, sensitivity, and inconclusive rates. This allows decision makers or the consumers of polygraph test data to calculate the probability of error and to choose what level of risk is acceptable. The

ESS was thought to be a more robust scoring technique since it allows the calculation of probability of errors and is simple to use.

Comparison of Systems

The following tables summarizes the results of a published study showing accuracy levels of the primary scoring systems used in TDA across the nation. The table highlights several important factors between the three systems. First, it shows that all of the systems tested have similar abilities to identify deception in during known field cases. It also shows inconclusive rates were also very similar and the ESS system had a higher rate of inconclusive when comparing the truthful test results only.

Table 4

Means, (Standard Deviations), and {95% Confidence Intervals} for Criterion Accuracy

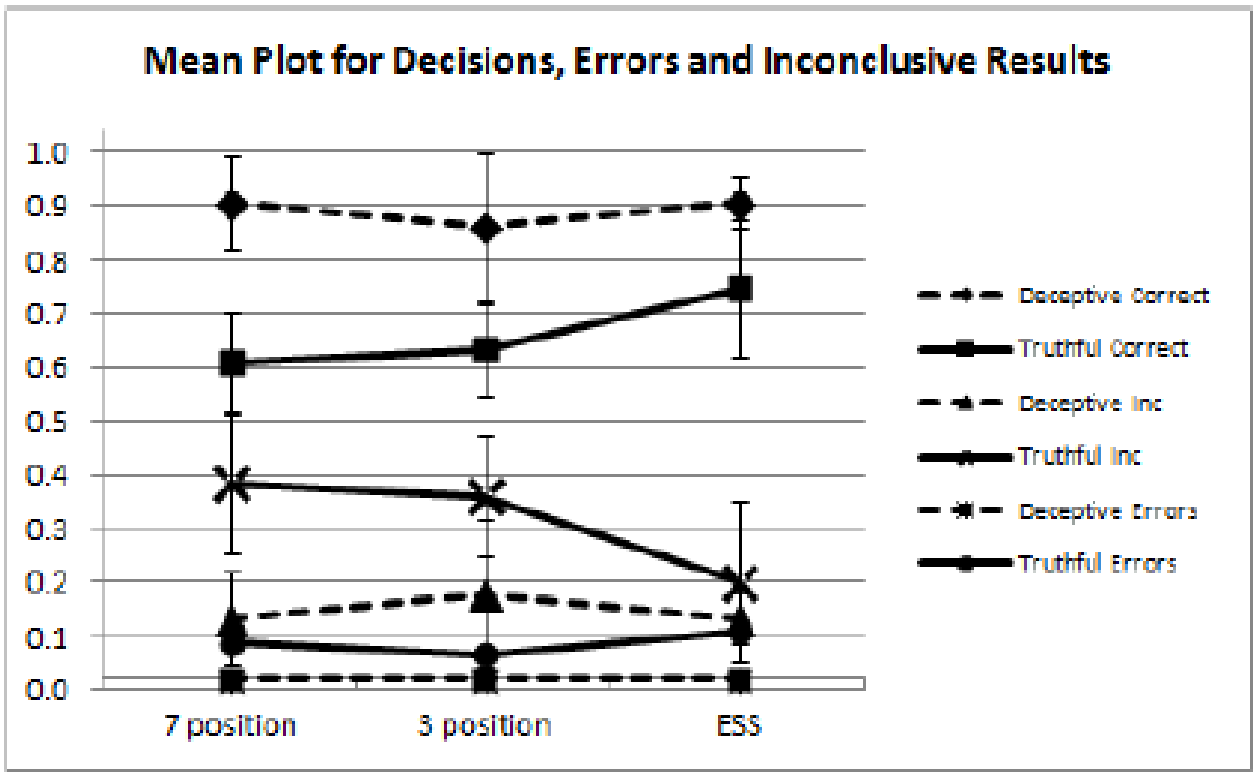
	3-Position	7-Position	ESS
Sensitivity	.886 (.087) {.716 to >.999}	.841 (.136) {.574 to >.999}	.886 (.045) {.797 to .975}
Specificity	.591 (.091) {.413 to .796}	.614 (.087) {.443 to .784}	.727 (.129) {.475 to .979}
Inc D	.114 (.087) {<.001 to .284}	.159 (.136) {<.001 to .426}	.114 (.045) {.025 to .203}
Inc T	.341 (.155) {.037 to .645}	.341 (.114) {.117 to .565}	.182 (.148) {<.001 to .473}
FN Errors	<.001 (<.001) {<.001 to <.001}	<.001 (<.001) {<.001 to <.001}	<.001 (<.001) {<.001 to <.001}
FP Errors	.068 (.087) {<.001 to .239}	.045 (.052) {<.001 to .148}	.091 (.074) {<.001 to .236}

(Robertson, 2012)

Table 5 takes the data shown in table 4 and displays the mean levels of correct truthful and deceptive outcomes, inconclusive results, and error rates on the scale of 0 through 1, with 1 indicating a high level of sensitivity and values near 0 indicating low rates of error.

Table 5

Mean Plot for Decisions, Errors, and Inconclusive Results



(Robertson, 2012)

Sensitivity as measured by Robertson (2012) was the test sensitivity to deception and specificity describes test specificity to the truth. The inconclusive D was the number of inconclusive tests that should have showed deception and inconclusive T tests are those that should have been evaluated as truthful exams. False negative (FN) errors describes an instance where a deceptive person is shown to be truthful and false positive (FP) errors are those where a truthful person is accused of deceptions.

The three-way contrast developed by Robertson (2012) encompasses all three of the primary systems being evaluated. Not only did this type of evaluation provide more freedom, it

was more powerful than several two-way analyses (Robertson, 2012). No statistically significant differences were observed in the accuracy of the three TDA systems that were tested. Although this study reported no false negative errors, this can be attributed to a number of potential factors including the small size of the study. These results should not be considered accurate as Krapohl (2006) reported a false negative error rate for ESS of 2.7%.

Conclusion

The 7-Position system, 3-Position system, and the Empirical Scoring System produce very similar results when used to evaluate polygraph data that has been obtained via a comparison question test format. Although each system may have slight advantages over one another in certain aspects of robustness, no one system clearly supplants any of the others. Research available fails to show that one particular system is the most accurate available. As a result, no recommendation can be made for the adoption of any particular manual polygraph scoring system as the national standard. Secondly, a recommendation to NCCA supporting inclusion of the ESS into their federal polygraph curriculum is not supported by existing research.

CHAPTER IV: RECOMMENDATIONS AND CONCLUSIONS

Future Research

Polygraph examinations will continue to be used by the federal government and other law enforcement agencies for determining the suitability and credibility of prospective employees for both reasons of character and national security. Specifically, there are numerous documented attempted infiltrations by either foreign government operatives or members belonging to groups that support or promote terrorism (Robertson, 2012). Additionally, the US government and others have a continual need to hire agents, as well as supplement their criminal investigations (Bondarenko, 2017). For these reasons, the necessity of standardizing numerical polygraph scoring across the profession supports the need for additional studies.

Based on this research, there are numerous recommendations that can be made regarding future studies. First, some of the studies reviewed here contained very small sample sizes, sometimes with as few as 15 confirmed cases. A study using a larger sample size may provide more statistical power and could show differences that these smaller studies may have missed. One example of this is the false negative error rate associated with the ESS system. One study has shown this system has a false negative error rate of 2.7 percent while a study with a small sample size showed no false positive rate of error (Robertson, 2012).

Secondly, a large sample size may help to review many other aspects of related research including criterion accuracy, decision accuracy and other issues. Because of this, the primary recommendation based on the included research is for the National Center of Credibility Assessment (NCCA) to conduct several larger studies to address some of the shortcomings that are present in previous research. Another issue that could possibly be addressed would include

having a better idea that the sample represents the population being studied. Third, other aspects that should be included within those studies are differing types of polygraph training, different testing formats used, the overall workload of examiners (background and investigative duties), and the amount of time required to use the different types of hand-scoring systems.

Lastly, since this paper supports the findings of numerous previous studies that the Empirical Scoring System shows promise as an additional TDA system, it is recommended that consideration be given to ESS by the federal polygraph community because it appears to offer examiners and the end users of polygraph data additional benefits. The two known benefits of using ESS would include the ability to calculate error rates and the ability for each program to alter decision rules to best fit individual program goals.

Clearly, no single study will contain a sample that perfectly represents the population as a whole. Because of this, the potential accuracy of any specific PDD examination scoring technique based on a single study should not be regarded as a definitive source. As with any field of research, additional study is warranted to help all stakeholders better understand and to confirm the capability of any scoring system.

Although a large study has not been conducted, the datasets that would allow for additional study do exist. The CIA, FBI, and USSS etc. all conduct polygraph examinations that include both lifestyle and national security components. The Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) as well as others conduct polygraphs limited to national security, but those can still provide insight when numerical scoring is the focus. As a result, the overarching recommendation would be to collect this data into a larger study where it can be used to determine statistically significant findings regarding the robustness of different numerical hand-scoring test data analysis systems.

Discussion

The impact of polygraph test results on an individual's future should not be taken lightly. Not only is this true in criminal matters, but equally as important in pre-employment screening scenarios. A failed polygraph test result can have a major impact on potential employees' future hopes of employment, across almost all levels of government. Although the results of polygraph examinations are not used in the courtroom, they are routinely used in hiring processes for law enforcement and intelligence agencies.

Although there is continued controversy, some of which that has received validation by the National Security Council, the use of polygraph is necessary to protect national security and combat criminal activity (Synnott, Dietzel, & Ioannou, 2015). The use of a polygraph instrument in the hands of a highly trained examiner is the best means the government has at detecting deception during those different types of investigations. Until a better way is found to differentiate between truth and deceptions, polygraph should and must continue to be leveraged to protect national security and the United States as a whole.

This paper supports the future adoption and standardization of any TDA system that exhibits the highest level of accuracy combined with the lowest possible error rates. That is not only right for accused criminals, but also for potential job applicants alike. The federal government should take the lead and do whatever necessary in order to recommend a TDA system as the national standard.

REFERENCES

Alder, K. (2002). A social history of untruth: Lie detection and trust in twentieth-century America. *Representations, 80*(1), 1–33.

Alder, K. (2007). America's two gadgets: Of bombs and polygraphs. *ISIS, 98*(1), 124-137. https://doi.org/10.1086/512836

Backster, C. (1969). Technique fundamentals of the tri-zone polygraph test. New York: Backster Research Foundation.

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

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal Of Neuroscience Methods*, *190*(1), 80-91.

Blackwell, N., & Department Of Defense Polygraph Inst Fort Mcclellan AL. (1998). *PolyScore 3.3 and Psychophysiological Detection of Deception Examiner Rates of Accuracy When Scoring Examinations from Actual Criminal Investigations*.

Bondarenko, V. (2017). Customs and Border Protection admitted to wasting more than $5 million on lie-detector test for ‘unsuitable’ job applicants. Retrieved on March 29, 2020 from: http://www.businessinsider.com/customs-border-protection-millions-lie-detector-tests-immigration-news-2017-8

Committee to Review the Scientific Evidence on the Polygraph, Board on Behavioral, Cognitive, and Sensory Sciences, & National Research Council (US). Committee on National

Statistics. (2003). The polygraph and lie detection. Washington, D.C.: National Academies Press.

Comparative Analysis Using DHS Screening Data, (2018).

Critchley, H. D. (2002). Electrodermal responses: What happens in the brain. *Neuroscientist*, 8, 132-142.

Davidson, W. (1979). Validity and Reliability of the Cardio Activity Monitor. *Polygraph, 08(2)*, 104-111.

Department of Defense (2006a). Test Data Analysis: DoDPI numerical evaluation scoring system. Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 10-03-2019.

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

Elaad, E. (2016). Extracting Critical Information from Group Members' Partial Knowledge Using the Searching Concealed Information Test. *Journal of Experimental Psychology: Applied, 22*(4), 500-509.

Fowles, D., Christie, M., Edelberg, R., Grings, W., Lykken, D., Venables, P. (1981) Publication recommendations for electrodermal measurements. *Psychophysiology, 18*(3):232–9.

Freedman, L., Scarpa-Scerbo, A., Dawson, M., Raine, A., McClure, W., Venables, P. (1994). The relationship of sweat gland count to electrodermal activity. *Psychophysiology, 31*:196–200.

Gougler, M., Nelson, R., Handler, M., Krapohl, D., Shaw, P., & Bierman, L. (2011). Meta-Analytic Survey of Criterion Accuracy of Validated Polygraph Techniques. *Polygraph, 40*(4), 1-118.

Grubin, D., & Madsen, L. (2005). Lie detection and the polygraph: A historical review. *Journal of Forensic Psychiatry & Psychology*, *16*(2), 357-369.

Gudjonsson, G. H. (2003). The psychology of interrogations and confessions: A handbook. Chichester: John Wiley & Sons.

Handler, M. (2015). Employee Polygraph Protection Act (EPPA). Retrieved from https://www.polygraph.org/employee-polygraph-protection-act-eppa-

Handler, M., Nelson, R., Goodson, W., & Hicks, M. (2010). Empirical Scoring System: A crosscultural replication and extension study of manual scoring and decision policies. *Pending publication*.

Harwell, E. (2000). A Comparison of 3- and 7-Position scoring scales with field examinations. *Polygraph, 29*(2), 195-197

Honts, C., Raskin, D., & Kircher, J. (1994). Mental and Physical Countermeasures Reduce the Accuracy of Polygraph Tests. *Journal of Applied Psychology, 79*(2), 252-259.

Horowitz, S., Kircher, J., Honts, C., & Raskin, D. (1997). The role of comparison questions in physiological detection of deception. *Psychophysiology, 34*(1), 108-115.

Keeler, L. (1930). Deception tests and the lie detector. International *Association for Identification Proceedings, 16,* 186–193

Kircher, J., Packard, T., Bell, B., & Bernhardt, P. (2010). Effects of prior demonstrations of polygraph accuracy on outcomes of probable-lie and directed-lie polygraph tests. *Polygraph, 39*(1), 22-67.

Kircher, J., & Raskin, D. C. (1981). Computerized decision-making in the detection of deception. *Psychophysiology,* 18, 204-205.

Kircher, J., & Raskin, D. C. (1988). Human versus Computerized Evaluations of Polygraph Data in a Laboratory Setting. *Journal of Applied Psychology, 73*(2), 291-302.

Krapohl, D. (1998). A comparison of 3-and 7-Position scoring scales with laboratory data. *Polygraph, 27*(3), 210-218.

Krapohl, D., (2006). Validated Polygraph Techniques. *Polygraph 35*(3), 149–55.

Krapohl, D. (2010). A Test of the Empirical Scoring System (ESS) with Two-Question Field Cases. Defense Intelligence Agency.

Krapohl, D., & Shaw, P. (2015). *Fundamentals of polygraph practice*. San Diego, CA: Elsevier.

Krapohl, D., Stern, B., & Bronkema, Y. (2009) Numerical Evaluation and Wise Decisions. *Polygraph, 38*(1), 57-71.

Krapohl, D., & Rosales, T. (2014). Decision Accuracy for the Relevant-Irrelevant Screening Test: a partial replication. *Polygraph, 43*(1), 20-29

Kristjansson, S., Kircher, J., Webb, A., Senter, S., & Dollins, A. (2006). Reliability and Validity of Physiological Measures for the Detection of Deception. Department of Defense Polygraph Institute. Fort Jackson, South Carolina

Larson, J. A., Haney, G. W., & Keeler, L. (1932). Lying and its detection: A study of deception and deception tests. Chicago, IL: University of Chicago Press

Matsuda, I., & Ogawa, T. (2011). Improved method for calculating the respiratory line length in the Concealed Information Test. *International journal of psychophysiology*, *81*(2), 65-71.

Merion, E., Krapohl, D.J., & Ashkenazi, T. (2008). An assessment of the Backster "Either-Or" Rule in polygraph scoring. *Polygraph, 37*(4), 240-249.

Moore, K. (2017). *Clinically oriented anatomy* (7th ed.). Philadelphia: LWW.

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

Nelson, R. & Handler, M. (2010). Empirical Scoring System: NPC Quick Reference. Lafayette Instruments: Lafayette, IN.

Nelson, R., Handler, M., Blalock, B., & Cushman, B. (2016). Blind scoring of confirmed federal You-Phase examinations by experienced and inexperienced examiners: Criterion validity with the Empirical Scoring System and the seven-position model. *Polygraph*, *45*(1).

Nelson, R., Handler, M., & Senter, S. (2012). Monte Carlo study of criterion validity of the directed lie screening test using the Empirical Scoring System and the Objective Scoring System version 3. *Polygraph*, *41*(3), 144-155.

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*(3), 185-215

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

Pollina, D., Dollins, A., Senter, S., Krapohl, D., & Ryan, A. (2004). Comparison of Polygraph Data Obtained From Individuals Involved in Mock Crimes and Actual Criminal Investigations. *Journal of Applied Psychology*, *89*(6), 1099-1105.

Raskin, D. C., & Hare, R. D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology*, *15*(2), 126-136.

Raskin, D. C., & Honts, C. R. (1987). The comparison question test. Handbook of polygraph testing, 1–47.

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). A study of the validity of polygraph examinations in criminal investigation (Grant No. 85-IJ-CX-0040). Salt Lake City: University of Utah, Department of Psychology.

Reid, J. E., & Inbau, F. E. (1977). Truth and deception: The polygraph (lie-detector) technique. Williams & Wilkins Company. Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law and Criminology, 37*, 542–547.

Robertson, B. (2012). *The Use of an Enhanced Polygraph Scoring Technique in Homeland Security: The Empirical Scoring System-Making a Difference*

Rovner, L. I., Raskin, D. C., & KIrcher, J. C. (1979). Effects of information and practice on detection of deception. *Psychophysiology, 16*(2), 983.

Slavkovic, A. (2018). Evaluating Polygraph Data.

Summers, W. G. (1936). Guilt distinguished from complicity. *Psychological Bulletin, 33*(9), 787.

Synnott, J., Dietzel, D., & Ioannou, M. (2015) A review of the polygraph: history, methodology and current status. *Crime Psychology Review, 1:1*, 59-83, DOI: 10.1080/23744006.2015.1060080

Topoglu, Y., Watson, J., Suri, R., Ayaz, H. (2020). Electrodermal Activity in Ambulatory Settings: A Narrative Review of Literature.

Use of Polygraph Examinations in the Department of Justice. (2006). Retrieved 5 March 2020, from https://oig.justice.gov/reports/plus/e0608/intro.htm

Warner, W. (2005). Polygraph testing: A utilitarian tool. *F.B.I. Law Enforcement Bulletin* 74, 4.

APPENDIX A

Definition of Terms

Psychophysiological Detection of Deception (PDD): The academic discipline that provides the student, the practitioner, and the researcher with the theoretical and applied psychological, physiological, and psychophysiological fundamentals for a thorough understanding of PDD tests, and the skills and qualifications for conducting PDD examinations.

PDD Examination: A process that encompasses all activities that take place between a PDD examiner and an examinee during a specific series of interactions. These interactions may include the pretest interview, the use of the polygraph instrument to collect physiological data from the examinee while presenting a series of tests, the test data analysis phase, and the posttest phase, which may include the interrogation of the examinee.

Relevant Question: A question that pertains directly to the matter under investigation or to the issue(s) for which the examinee is being tested.

Comparison Question: A question that is designed to produce a physiological response. The physiological responses of the comparison questions are compared to the physiological responses of the relevant questions. The probable and directed lie are the two types of comparison questions utilized within the federal government

Probable Lie Comparison (PLC) Question: This question is designed to be a probable-lie for the examinee. The PLC question should be similar in nature but unrelated by time, place or category to the specific issue. However, in screening examinations the PLC can be related to the issue(s) as long as the screening comparison question establishes a dichotomy between the relevant and comparison issues. A comparison question should be broad in scope and time so

that it captures as many of the examinees past life experiences as possible. The physiological responses to the PLC are compared to the responses of the designated relevant questions. The exclusionary and screening comparison questions are the two types of PLC questions used within the federal government.

Directed Lie Comparison (DLC) Question: A specialized comparison question addressing a minor transgression to which most people will readily admit. Upon acknowledging having committed such a transgression, the examinee is directed to lie when asked that question on the test.

Irrelevant Question: A question that is designed to be non-emotion evoking and unrelated to the issue being tested.

Personnel Security Screening (PSS) PDD Examination: A PDD screening examination conducted to aid in determining an individual's eligibility for initial or continued access to designated programs or information, or an examination conducted to aid in determining an individual's eligibility for initial access to sensitive law enforcement positions.

Specific Issue PDD Examination: A PDD examination conducted to resolve a specific issue, e.g., criminal, espionage, sabotage, or source validation.

Deception Indicated (DI): An opinion which indicates that an analysis of the PDD charts revealed the physiological responses to the relevant question(s) were indicative of deception.

No Deception Indicated (NDI): An opinion that indicates that an analysis of the PDD charts revealed the physiological responses to the relevant question(s) were not indicative of deception.

No Opinion (NO): An evaluation which indicates the examiner cannot render an opinion based upon the physiological data on the charts.

No Significant Response (NSR): This opinion indicates that the analysis of the PDD charts revealed no consistent, significant, timely, physiological responses to the relevant questions in personnel screening, source validation, or POT tests.

Significant Response (SR): An opinion which indicates that the analysis of the PDD charts revealed consistent, significant, timely physiological responses to the relevant questions in personnel screening, source validation, or POT tests.

Test Data Analysis: The analysis of the psychophysiological responses recorded on the PDD charts. Only data that is timely with the applied stimulus and free of artifacts and unwanted noise on the signal can be evaluated. (Department of Defense, 2006b)