The Use of an Enhanced Polygraph Scoring Technique in Homeland Security: The Empirical Scoring System—Making a Difference

Bruce P. Robertson[1]

Abstract

This thesis studies the polygraph Empirical Scoring System (ESS) to determine its potential use in homeland security and the war on terror. The research based its analysis on raw data previously collected by other researchers, who removed identifications from the data and subsequently provided it for study here. The results are described in regards to criterion accuracy; diagnostic capability; proportions of correct, errors, and inconclusive results; and the difference in scoring accuracy based upon participant employment and experience. Twelve scorers in three cohorts scored 22 You-Phase examinations taken from the Department of Defense–confirmed archives. One cohort used the three-position test data analysis (TDA) system, another cohort used the seven-position TDA system, and the final cohort used the ESS TDA system. All TDA systems proved equally capable of diagnostic ability. ANOVAs showed no significant differences between the distributions of ESS and transformed scores. No significant differences were found in decision accuracy with correct, inconclusives, errors rates for ESS scores, and those from the other two TDA systems. That ESS can complement other current hand-score TDA systems is suggested. However, that it could supplant other TDA systems is not confirmable by this study. Further study is recommended.

Introduction

Physiology has been used in the United States in the detection of deception since World War I, when the government commissioned Dr. William Marston[2] to devise a technique to question prisoners of war (Alder, 2007). Intelligence officials from the National Research Council sponsored Marston's research. In his first real-world case, Marston used his techniques to attempt to identify the culprit in the theft of a military codebook from the United States Surgeon General's office. Although he narrowed the field of suspects to one, there is no record that the identified man was ever charged or in fact had committed the theft (Adler, 2007). Method more than instrumentation was Marston's contribution to lie detection. He believed that, by monitoring changes in systolic blood pressure, verbal deception could be detected. As described by Ball and Gillespie on their website:

> He used a standard blood pressure cuff, or sphygmomanometer, to take measurements of systolic blood pressure during interrogation. This was the first time anyone used any kind of an instrument to detect truthfulness or deception. His method was simple. Take and record the subject's blood pressure, release the cuff. Ask the subject a question. Take and record the subject's

[1] The author is the Chief of Police for Centerville, Ohio. This paper was submitted in partial fulfillment for his Masters degree from the US Naval Postgraduate School in 2012, and has been formatted for republication in *Polygraph*. The original paper can be downloaded at dtic.mil. The opinions expressed in this paper do not necessarily represent those of the American Polygraph Association.

[2] Dr. William Marston was a Harvard psychologist who is likely better known for the creation of the comic book character "Wonder Woman" under the nom de plume "Charles Moulton."

> blood pressure once again to identify any changes. He called this the "discontinuous method" of detecting deception. (Ball & Gillespie Polygraph, n.d.)[3]

The polygraph, with this long and controversial history, has been used at the federal, state, and local levels for a variety of purposes ever since. These uses include criminal cases, pre-employment screening, informant and witness testing, and counterintelligence purposes (Warner, 2005). There are 26 federal polygraph programs spread across nine federal agencies (see Table 1 for a listing of the polygraph programs), as well as numerous state and local law enforcement agencies.

Table 1. Federal Agencies That Utilize the Polygraph[4]

Department of Defense	Non-Department of Defense
Air Force Office of Special Investigations	Alcohol, Tobacco and Firearms
Army Intelligence Polygraph Program[5]	Bureau of Prisons/Office of Internal Affairs
Defense Criminal Investigative Service	Customs and Border Protection/Internal Affairs
Defense Intelligence Agency	Coast Guard Investigative Service
Naval Criminal Investigative Service	Central Intelligence Agency
National Geospatial-Intelligence Agency	Drug Enforcement Administration
National Reconnaissance Office	U.S. Department of Energy
National Security Agency	Federal Bureau of Investigation
U.S. Army Criminal Investigation Command	Food and Drug Administration
	Homeland Security Investigations[6]

[3] In a scholarly article on the history of the polygraph, Paul Trovillo notes that Angelo Mosso, an Italian psychologist who studied under Cesare Lombroso, first experimented with a plethysmograph to study the effects of fear on human blood pressure. These experiments, as well as several that came later, were viewed as instrumental in the early study of the polygraph. In 1895, Lombroso, an Italian physician, psychiatrist, and criminologist, modified a medical instrument known as a hydrosphygmograph (similar to a modern cardiophymograph) to measure the blood pressure and pulse rate of a criminal suspect under police interrogation. This is believed to be the first application of a mechanical instrument for lie detection (Trovillo, 1972).

[4] As of February 3, 2012.

[5] Formerly the United States Army Intelligence and Security Command (INSCOM).

[6] Formerly Immigration and Customs Enforcement (ICE).

Table 1. Federal Agencies That Utilize the Polygraph (cont.)

	Internal Revenue Service-Criminal Investigation
	Transportation Security Administration
	U.S. Postal Inspection Service
	U.S. Postal Inspection Service, Office of Inspector General
	United States Secret Service

The controversy over polygraph validity and reliability is ongoing, but the utility of the polygraph to obtain information is widely acknowledged (Warner, 2005). In homeland security and the war on terror, the polygraph has many applications. Specifically, it has been used by intelligence and other federal agencies for counterintelligence and espionage purposes. Many agencies use it as part of ongoing security screening programs for current employees. The Central Intelligence Agency's (C.I.A.) Aldrich Ames case and the Department of Energy's Wen Ho Lee[7] case are just two controversial examples of polygraph use in espionage investigations at the federal level. These two cases exemplify why scoring techniques are so important to the field and why poor technique or diagnostics or lack of interrater reliability can be detrimental to national security. The common public perception is that Ames passed his polygraphs (Alder, 2007; Pentagon's intelligence arm, 2008) while the polygraph was partially responsible for the bungling of the Lee investigation (Hoffman & Stober, 2001; Alder, 2007; Wen Ho Lee's Problematic Polygraph, 2000). It is myth that the polygraph's alleged failure allowed the two men to continue their deception.[8] These cases raised questions about the very foundation on which the government bases its use of the polygraph for national security purposes. A brief look at each case will demonstrate some of the issues of test data analysis, the component of the polygraph process that this research studies.

Ames, a Central Intelligence Agency Directorate of Operations officer, was arrested in 1994 for selling information to the Soviet Union. According to publications, Ames had spied for the KGB for nine years, and his duplicity had resulted in the death of at least ten agents who had spied for the C.I.A. in the Soviet Union (Earley, 1997; C.I.A., n.d.). In 1994, Dan Glickman, the House Intelligence Committee chairman, noted that the Federal Bureau of Investigation had concluded that Ames did not pass either of two tests (Kleiner, 2002). Then-C.I.A. Director James Woolsey in 1994 revealed that the F.B.I. had not properly investigated Ames's two failed polygraphs (Kleiner, 2002).

Wen Ho Lee, a naturalized United States citizen, became suspect as a Chinese spy in 1995, after his employer, the Department of Energy (DOE), deduced that China had stolen classified nuclear weapons

[7] It is of note that the National Academy of Sciences believed the Lee case so important to the government's reliance on the polygraph that it devoted an appendix to the case in its report (*The polygraph and lie detection*, 2003) and that eNotes, a popular research site for students and teachers, uses it as its case study for polygraph on its website (Lerner & Lerner, 2006).

[8] This statement is based on personal interviews with primary sources who cannot be identified due to security concerns. These personal conversations have taken place during the 25-plus-year polygraph career of the author.

designs that allowed the country to develop a miniaturized nuclear warhead (Wen Ho Lee Case Study, 2008). Lee had been employed at the DOE's Los Alamos National Laboratory in New Mexico since 1978 and later became a nuclear weapons scientist at the laboratory. During the 1980s and 1990s Lee had numerous contacts with Chinese officials and scientists, some on official business and others while attending parties or conferences (Hoffman & Stober, 2001). As part of his employment, Lee was subject to periodic polygraph examinations: one in 1984, one in 1998, and another in 1999 (*Polygraph and lie detection*, 2003). However, the results of these polygraphs are in dispute (*Polygraph and lie detection*, 2003; Wen Ho Lee's Problematic Polygraph, 2000). More specifically, it is the disagreement between the opinions rendered by the original polygraphist and other polygraphists who later reviewed the polygraph charts as to whether or not Lee was truthful that creates the issues of concern.[9] The interpretation and scoring of polygraph charts is the focal point of this thesis.

The polygraph is used by federal, state, and local governments in determining the credibility and suitability of prospective employees who potentially will have a role in homeland security and/or the war on terror. The author has primary-source information that polygraph pre-employment screening in a major city law enforcement agency uncovered two attempted infiltrations, one by a Chinese operative and the second by a member of Al-Qaeda.[10] In the case of the Chinese operative, the agent was to gain employment at a law enforcement agency and work there long enough to establish a record of credibility in order to later become an employee of a federal law enforcement agency. In the case of the Al-Qaeda affiliated applicant, the effort was just an attempt to infiltrate law enforcement in a major city, one with a large Muslim population.

The polygraph has been used in Guantanamo Bay, Kandahar, Bagram, and other front-line combat theatres. In September 2003, the Air Force Office of Special Investigations (AFOSI) deployed its first full-time polygraphist to Baghdad (Collins, 2004, p.1). Prior to this, the Air Force had deployed polygraphists on temporary duty (TDY). The (then-) polygraph program manager, Special Agent Pat Muller, was quoted as saying, "The polygraph exams we have administered over there have been some of the most critical and important work we have ever done in this program" (Collins, 2004, p. 1). The scope of examinations in the theatre of war includes vetting coalition force members, determining the veracity of prisoners and informants on whose information tactical operations are initiated, and assisting in the conduct of criminal investigations (Collins, 2004, p. 1).

Problem Statement

Information provided to decision makers should be as accurate, trustworthy, and robust as possible, and it is clear that the polygraph plays an important role in achieving these requirements. Each day decision makers in federal, state, and local governments rely on the results of polygraph examinations to make their decisions. In its 2002 Polygraph Program Annual Report to Congress (Department of Defense, 2003), the Department of Defense (DoD) reported that it had conducted 11,566 polygraph examinations.[11]

[9] What does not seem to be at issue is that Lee illegally removed huge amounts of classified nuclear information from the laboratory, estimated at over 400,000 pages, and that once removed, its final destination(s) have never been learned (Shelby, 2001).

[10] Due to the classification, sensitivity, and state civil-service rules, neither the name of the agency nor the minute details can be divulged.

[11] This is the final report made to Congress, since Congress relieved the DoD of its reporting responsibilities after fiscal year 2002. No current figures are available as to the number of examinations currently conducted by the DoD. In 1991, Congress authorized the DoD to conduct no more than 5,000 CSP examinations annually. However, this quota was lifted in 2005, and there is currently no cap on CSP examinations. The figure of 8,512 includes those conducted by the DoD for non-DoD federal agencies. It is also noted that these numbers include only the DoD and not the National Security Agency or those conducted under the authority of the director of Central Intelligence

The possible results of a polygraph examination are: "no deception indicated" (passed), "deception indicated" (failed) or "inconclusive" (the tracings were such that no opinion can be rendered). These judgments are rendered using one of several scoring mechanisms. None of these manual scoring systems in common use will deliver error estimates except ESS. That is to say, there is no current scoring mechanism that allows the polygraphist or the consumer to compare a calculated probability of error to a stated tolerance for error (Handler et al., 2010). The p-value maps the scores over a probability distribution such that the consumer can estimate the error likelihood of a decision based on the scores. These error estimates allow the consumer to take a more informed value judgment about tolerance for risk or error. Other current scoring systems in use do not have the same empirical level of decision accuracy as ESS (Handler et al., 2010). ESS provides accuracy profiles to include the total proportion of correct, inconclusive, deceptive, truthful, sensitivity, specificity, false negative errors (liars called non-deceptive), and false positive errors (truthful called deceptive).

This has been the state of the profession since the early use of the polygraph, much to the derision of its critics. The employment of an empirically based scoring mechanism would allow polygraphists to render an opinion based upon confidence in a scientifically derived result. The questions therefore become: does the scoring mechanism that provides that p-value have at least the same or better accuracy profiles as current scoring mechanisms; how can it be applied, and would it be accepted?

Research Question

The broad question under consideration is whether the accuracy profiles associated with various scoring techniques should have an impact on the technique chosen in the homeland security arena. Additional questions to support this analysis will include: 1) Are there differences in the effectiveness of the three-position, seven-position, and ESS test data analysis (TDA, chart interpretation) models at extracting diagnostic information from the raw data, as reflected by the distributions of numerical scores? 2) Are there significant differences in criterion accuracy for the three-position, seven-position, and ESS TDA models? 3) What is the effect on accuracy of transforming three-position and seven-position scores to ESS scores? 4) How accurate are the combined three-position, seven-position, and ESS results? How accurate are the combined results when all scores are transformed to ESS scores? Is the difference significant? 5) Are there differences in accuracy that can be attributed to experience? Does more experience result in increased accuracy? 6) Does accuracy vary with the examiner's type of employment? Are there differences between private examiners and those who work for government (law enforcement/federal government) agencies?

Literature Review

The polygraph has demonstrated an important role in homeland security and the war on terror. This role has included the screening of personnel within many federal, state, and local agencies across the United States to assist in ensuring that prospective hires do not have an illicit motive for joining the ranks. It is not only important to understand that criminals and terrorists alike have attempted and been successful in acts that threaten homeland security, but enemies of the nation have the intent to spy and/or recruit potential agents for the purposes of espionage within our intelligence agencies and throughout other levels of government. The polygraph was used in World War I in counter-intelligence operations. It gained greater and more specific use in Korea (Alder, 2007). Since then, it has been used to assist decision makers in taking strategic and tactical decisions that directly protected American assets and lives as well as those of our allies. This review will identify literature about polygraph scoring techniques that are currently relevant to the topic, as well as those that will enhance the reader's understanding of the field of the polygraph.

This literature review will address three areas related to the polygraph and hand-scoring techniques. The first section will give a brief overview of polygraph research. The second section will provide an overview of types of testing techniques. Finally, the third

section will discuss research related to scoring techniques.

Polygraph Research

Polygraph expert and researcher Stuart M. Senter claims that polygraph examination is an inimitable field: "Polygraph examiners are trained to accomplish a task that, in the mind of the public, should only be made possible through rapid advances in seemingly futuristic technological equipment or through the weaving of mystical powers thought to be proffered by wizards and magicians." In other words, Senter is implying that many consider polygraph nothing but a magic trick, subject to ridicule and derision (Senter, 2008). Senter goes on to note that providing a more pragmatic view of the polygraph will be accomplished through increasing the body of knowledge about the field. To date, the research has focused on applied research. That is, it focuses on real-world problems and tends to ignore theoretical knowledge. However, the basic foundations of polygraph principles have been ignored. There is little work on the understanding of factors that look at the diagnostic value of the polygraph (Senter, 2008, p. 278).

The National Research Council points out that there must be a solid theoretical base to have confidence in polygraph tests, lest erroneous results in populations such as "spies and terrorists" fail national security (*Polygraph and lie detection*, 2003, p. 92). However, the field has not made proper use of theoretical systems about the processes that underlie the measurements taken by the polygraph (*Polygraph and lie detection*, 2003, p. 93). Further, the research on the concept of decision thresholds (which are part of scoring techniques) has largely been ignored in polygraph research.

The consensus is that, although improving, in order to bring the polygraph into the realm of a recognized science, robust research must continue to be pursued.

Testing Techniques

This section is not an exhaustive overview of all testing techniques in use in the field of polygraph examination. It is a literature review of sources pertaining only to the most common techniques currently being utilized. Donald Krapohl and Shirley Sturm, in their 2002 article in *Polygraph* identify a number of testing techniques. The Air Force Modified General Question Test is a single-issue, multiple-issue, or multi-facet technique (Krapohl & Sturm, 2002). The Comparison Question Technique is a term applied to a number of test formats that use probable- or directed-lie test questions. A Concealed Information Test is a type of test that involves a series of tests in which one critical item is used in each series. The intent of the test is to determine the person's knowledge of the particular item. A Counterintelligence-Scope Polygraph (CSP) is a type of test given to federal government employees who have access to sensitive security information. The CSP is designed to "detect and deter espionage, security breaches, sabotage, or other acts against the government" (Krapohl & Sturm, 2002, p. 172). A test format that is widely used in the field is known as the Modified General Question Test (MGQT). The MGQT consists of more relevant questions than comparison questions. It does not use what is known as a "symptomatic question."[12] A Modified Relevant/Irrelevant Technique is a specific-issue test that uses situational comparison questions, which are then compared to the relevant questions. The relevant/irrelevant technique is a family of test formats that forgo the use of a traditional comparison question. They are most widely used in screening tests. U.S. government agencies use a test known as the Test for Espionage and Sabotage, which is a multi-issue screening test typically used with

[12] A "symptomatic question" is a question used to identify whether or not an examinee is fearful that the polygraphists will ask an unreviewed question embracing an outside issue that is bothering the examinee. This mistrust of the examiner will putatively dampen the examinee's responses to other test questions. Symptomatic questions are widely used, though the trend in the research is that there is no meaningful effect (Krapohl & Sturm, 2002).

government employees who have access to sensitive information or programs/projects. The Utah Technique is a technique that uses modules of questions that consist of a comparison, relevant, and irrelevant question. The You Phase is a single-issue test in which the relevant question is slightly varied throughout the test. It is a highly focused test. The Zone Comparison Test (ZCT) uses three zones that refer to categories of questions (relevant, comparison, and symptomatic) that then compare two of the zones (relevant and comparison) to determine whether the examinee was truthful or deceptive. It is designed to "focus their attention to specific zone question(s). It is the first modern polygraph technique to which numerical analysis was widely applied" (Krapohl & Sturm, 2002).

Scoring Techniques

The global evaluation technique is one in which the polygraphist visually inspects the charts to determine whether there is a stronger response to the relevant questions. It is most commonly used to score the Relevant/Irrelevant Technique (RI Technique). The NRC, as well as Krapohl and Dollins, notes that there is a lack of standardization to the scoring technique and that it has numerous idiosyncrasies (*Polygraph and lie detection*, 2003; Krapohl & Dollins, 2003). Literature on this technique is scant, and its general use has declined.

The technique favored by most current polygraphists is numerical scoring in its several variations. The introduction of numerical scoring for the Comparison Question Technique is attributed to Cleve Backster, a well-known school director and instructor in modern polygraph techniques (Weaver, 1980). He introduced the seven-position scoring TDA system. The scale assigns scores ranging between +3 and -3 to the respective questions and their "comparison" questions. Weaver notes that the scoring technique was first developed by Backster to assist students in chart analysis in classroom settings. In later research conducted by the University of Utah, it was concluded that numerical scoring had higher rates of accuracy and reliability than other scoring techniques (Raskin, Barland, & Podlesny, 1978), and it became the benchmark for the profession. The scoring system has evolved to include a three-position TDA system. This scoring system is now in wide use by polygraphists.

Krapohl and Dollins undertook what they described as a rudimentary investigation of the three primary scoring rule systems that can be applied to these numerical scoring techniques (Krapohl & Dollins, 2003). The three scoring systems are known as the Utah, the Backster, and the federal scoring systems. These scoring systems have three common components: scoring rules, computation rules, and decision rules (cut scores) (Krapohl & Dollins, 2003, p. 150). It is important to understand these three terms as used in the literature as they will be explored further as part of this research. Scoring rules are those that relate to the choice of tracing features in the charts, rejection of artifacts, and the choice of how question pairs are compared and numbers assigned to the scheme. The weight and how the numbers are combined describe the computation rules. Decision rules, otherwise known as cut scores, govern the relationship between the computation rules and the polygraphist's choice of a decision (opinion), which will either be Deception Indicated (DI), No Deception Indicated (NDI) or inconclusive (INC) (Krapohl & Dollins, 2003).

Decision rules predominated in conclusions reached by the NRC and Krapohl, Stern and Bronkema (*Polygraph and lie detection*, 2003; Krapohl, Stern, & Bronkema, 2009). Specifically, each came to the conclusion that risk tolerance, and the corresponding decision rules, should be set by the consumer of polygraph results. That is, this decision should not be left to the polygraphists\ but to the consumer of the results, who ultimately decides what risk can be accepted in the decision making process. In short, the determination of decision rules is a policy decision and will come into play later in the discussions of this research.

Two things become apparent in the literature: Those who speak to the topic agree on the paucity of research into the polygraph, and some note that the research concerning scoring techniques is even rarer. Secondly, the

research into hand-scoring techniques looks into many things. Prior research includes accuracy and reliability of the scoring technique and the relative simplicity or lack of it within the respective technique and interrater reliability. What prior research lacks is the incorporation of the study of normative data (Handler et al., 2010).

Another scoring technique—the topic of this research—is the Empirical Scoring System (ESS). This scoring system was first described by Krapohl, Nelson, and Handler in 2008 (Krapohl, Nelson, & Handler, 2008). The development and research conducted on ESS allowed for the first time in the development of a polygraph hand-scored technique the application of p-values and normative data. It is profound in its simplicity, and based on associated p-value tables in regard to specificity, sensitivity, and inconclusive rate, the decision maker or policy setter can compare the probability of error and choose the error rate that best fits into his schema for risk aversion. It is because of this unique ability, in conjunction with the simplicity of its use, that ESS may prove to be the most robust scoring technique and capable of protecting American lives and assets at home and in the field of combat.

Hypotheses or Tentative Solutions

The polygraph is used in many circumstances for the purposes of national security, as well as law enforcement and security issues at the state and local levels. Its use in combat zones as well as the rear areas in theatres of war is documented. It has proven to be an extremely useful tool by assisting decision makers in the field to make both strategic and tactical decisions. The claim is that, by providing polygraph experts with a simpler hand-scoring technique, based on empirical data to which probability values have been determined, they in turn can provide these decision makers with a more informative answer to the questions at hand. In the combat arena, those questions can revolve around whether or not to undertake a tactical operation based on the word of an informant, collaborator, or captured enemy combatant. Such decisions involve great risks to life and limb, and the decision makers must be given the best tools available to make them. In other homeland security concerns, they can involve the credibility of informants, witnesses, accused or suspected criminals, spies, and other ne'er-do-wells.

Evidence to support this claim can be found in the review conducted by the National Research Council (NRC). The NRC notes that

> decision scientists and policy advisers have worked to develop systematic methods for resolving hard decision problems that arise in business, medicine and public policy. These methods are used explicitly in many scientific articles, and they are used implicitly in practical advice, where the goal is to get decision makers to think systematically before acting. (*Polygraph and Lie Detection*, 2003, p. 358)

The history of the polygraph is such that the lack of a sound scientific basis, in the minds of some, has led to the dismantling of various polygraph programs,[13] caused decision makers to be reluctant to rely on it—even in the absence of alternatives—and caused much conversation in the halls of Congress, state houses, and local government buildings as to its usefulness. It is a proven tool in the war on terror and national security. The Empirical Scoring System is the simplest hand-scoring technique to have empirical and scientific support as its foundation.

Significance of Research

Literature

There is a dearth of literature on scientific and empirically based hand-scoring techniques in the field of polygraph examination, particularly the impact of the techniques used on the robustness of

[13] No polygraph programs have been dismantled at the federal level, and new federal programs have, in fact, been added since the 2003 NRC report. However, legislative decisions and court rules have impacted or outlawed polygraph programs at the state and local levels.

decisions taken by those who rely on the polygraph to assist them in their decision making process. The Empirical Scoring System is one of the first and simplest hand-scoring techniques with intent to anchor TDA on empirical evidence and scientific study (Handler et al., 2010). This research should not only impact the use of the polygraph as it relates to national security, homeland security, and the war on terror, but it should further the scientific advancement in the polygraph community as a whole.

Future Research Efforts

This research will reinforce the concept that a solid scientific basis for the polygraph will enhance its use and make it more readily defensible. The National Research Council (NRC) has stated that no lie detection technique has been shown to outperform the polygraph and none shows any promise in the near term (*Polygraph and lie detection*, 2003, p. 173). However, it also notes that past efforts at polygraph research have not laid a sound foundation of scientific knowledge in the field (*Polygraph and lie detection*, 2003, p. 213). On page 221 of its review (*Polygraph and lie detection*, 2003), the NRC goes on to say that the detection of deception and information withholding is important to national security and that "government agencies will continue to seek accurate ways to detect deception by criminals, spies, terrorists, and others who threaten public safety and security interests." This thesis is just one small part of this effort, and it is hoped that it encourages others in the field and those who are consumers of its product to engage in further scientific study, particularly as it relates to security on the national, state, and local levels.

Consumers

The immediate consumers of this research are the Department of Defense and its various military branches, as well as all federal agencies that have polygraph programs in place as part of their national and internal security interests. Further, all state and local law enforcement and criminal justice agencies who rely on polygraph results as part of their decision-making process should find this research useful. It is anticipated that the three national polygraph associations—the American Association of Police Polygraphists, the American Polygraph Association, and the National Polygraph Association—will utilize this research in the training and education of their respective members.

Homeland Security Practitioners and Leaders Nationally

This research should be of interest to many federal program managers within DHS and various federal agencies outside DHS, both those who use the polygraph and others who may not for various reasons. As this is just one small step in an effort to roll a component of lie detection onto a sound scientific basis, it can be anticipated that those who have been reluctant to utilize the polygraph, or perhaps even those who have been detractors of the field, might be encouraged and convinced to reconsider their positions.

Method

The present research based its analysis on raw data previously collected by other researchers instrumental in the development of the Empirical Scoring System (ESS), who removed identifiers from the data and subsequently provided it for study here.

Data was obtained from three groups (Cohorts 1, 2, and 3) of four scorers each. These participants were randomly grouped volunteers from a group of 300 students trained in the Empirical Scoring System as part of a training seminar hosted by the American Association of Police Polygraphists[14] in Cambridge, Massachusetts, on March 28, 2011. Cohort #1 scored the sample examinations using the Empirical Scoring

[14] The American Association of Police Polygraphists is the largest law enforcement polygraph association in the world. The author is both a past and current president.

System (Appendix A).[15] Cohort #2 scored the examinations with the three-position Test Data Analysis (TDA) system (DACA, 2006) (Appendix B), and Cohort #3 scored them using the seven-position TDA system (DACA, 2006) (Appendix C).

The Empirical Scoring System is an evidence-based numerical hand-scoring technique used for test data analysis of polygraph charts obtained from comparison question tests (Nelson et al., 2012). The ESS system utilizes a three-position scale of +, 0, or - and relies on the *bigger-is-better rule*;[16] scores are assigned when the scorer visually observes a difference in reaction strength between relevant and comparison questions (Nelson et al., 2012). A positive score (+) is assigned when there is a larger response to a comparison question, and a negative (-) score is assigned when there is a larger response to a relevant question. In typical comparison-question test formats, relative questions are normally compared to comparison questions (Nelson et al., 2012).

In "Terminology Reference for the Science of Psychophysiological Detection of Deception"[17] (Krapohl & Sturm, 2002), the seven- and three-position TDA systems are defined as follows:

7-position scale

System of assigning values to individual physiological responses in PDD, based on differential responding to relevant and comparison questions. The values in 7-position scoring are whole numbers between -3 and +3. By convention, negative values represent greater responding to relevant questions, while positive values indicate greater responses to comparison questions. A zero usually indicates equal or no reactions to the relevant and comparison questions, or that the spot does not meet minimum standards for interpretation. The assigned numbers are summed across all three PDD parameters for each question for all spots and all charts. There are thresholds for determinations of truthfulness or deception, with an inconclusive region separating them. In the PDD literature, the 7-position scale is sometimes referred to as a semi-objective scoring system. There are three major versions of the 7-position scoring system: Backster, Utah, and DoDPI. See: Bell, Raskin, Honts, & Kircher (1999); Swinford (1999); Weaver (1985).

3-position scale

Abbreviated form of the 7-position scale for PDD test data analysis. The major difference is that the range of values for each comparison is from -1 to +1, rather than the range of -3 to +3 in the 7-position scoring system. See: Capps & Ansley (1992); Krapohl (1998); Van Herk (1990).

The analysis method applied to the research questions was an analysis of variance (ANOVA), which will be further described for each research question.[18]

[15] Editor's Note: None of the listed appendices are reprinted in this publication due to space considerations. Readers interested in them can download the original document at the Defense Intelligence Technical Information website at dtic.mil.

[16] The instructions for the rule are simple: if you can see it, point to it, and support that the reaction is bigger, then you score it. If you can't point to it and support it, then do not assign a score.

[17] "Psychophysiological detection of deception" (PDD) is a term used primarily by the federal government and is interchangeable with the terms "polygraph" and "lie detector."

[18] Special gratitude is expressed to Raymond Nelson for his computational assistance.

Results and Analysis

Each study participant provided a demographic data form (see Appendix D). This demographic data included age and experience as a polygraphist, as well as gender.

The average age was 54, with a standard deviation of three. The maximum age was 65, and the minimum age was 37. The median age was 58. Ages do not appear normally distributed.

There were ten males and two females. Females n=2 is too small for analysis. Compared to groups of equal size, differences in the group size is significant. Z=9.334 (p<.001) Test of Proportions. Gender was not evaluated as an independent variable in the remainder of the analysis.

The average years of experience were 15, with a standard deviation of three. Median experience was 14 years, and the mode was also 14 years. The maximum years of experience were 33, and the minimum was three. Proximity of the mean, median, and mode indicated no increased concerns regarding the normality of the distribution of participant ages.

The participants in the study included four private examiners, seven law enforcement examiners, and one federal examiner.

Additional data collected on the hand-score sheets were the individual scores assigned by the participant to the two relevant questions on the three charts of each examination in the study sample. A score was assigned, according to the structured rubric for each scoring system, for the tracings of each of these sensors: pneumograph,[19] electrodermal (EDA),[20] and cardiograph.[21] Subtotal scores were calculated for each of the relevant questions, and a grand total was calculated for the test as a whole. Scores were then interpreted using structured decision rules, according to the requirements of each scoring method, to make categorical determinations as to no deception indicated (NDI),[22] deception indicated (DI),[23] or inconclusive (INC).[24] Inconclusive is sometimes referred to as "no opinion" or "indefinite." Each participant then rendered his personal confidence level in the opinion rendered (see Appendix E).

[19] The pneumograph sensors, one tube placed around the abdomen and another around the thorax, record respiration data. Features included in the manual scoring model pertain primarily to suppression or reduction of respiration activity.

[20] Changes in the electrical properties of the skin (exosomatic and endosomatic) typically measured by placement of electrodes on the central pad of skin of two fingers. This term superseded the term "galvanic skin response" (GSR), which can still occasionally be found in the older literature.

[21] A term for recording heart activity, typically done by placement of a blood pressure cuff on an arm, which then measures pulse wave and changes in relative arterial blood pressure. In this context it is more correctly called sphygmograph or plethysmography (Krapohl & Sturm, 2002).

[22] No deception indicated, in layman's terms, means that it is the polygraphist's opinion that the person is truthful as to the matter at hand.

[23] Deception indicated, in layman's terms, means that it is the polygraphist's opinion that the person is not truthful (lying) to the matter at hand.

[24] Inconclusive, in layman's terms, means that the polygraphist has no opinion as to whether or not the person is truthful or lying to the matter at hand. It is typical that "no opinion" is rendered when the diagnostic quality of the tracings is such that they cannot be analyzed. It is the author's experience that those within the field of polygraph scoring do not consider "no opinion" as an error and that in many cases with subsequent testing (sometimes called a "reexamination") a definitive opinion can be rendered.

However, it is duly noted that some outside the profession consider "no opinion" to be an error, and in research this dissenting opinion is sometimes taken into account.

Research Question #1

Results

Do differences exist in the effectiveness of the three-position, seven-position, and ESS test data analysis systems at extracting diagnostic information from the raw data?[25]

It is important to understand that the end result of any polygraph examination, whether for event-specific criminal investigations, security screening, law enforcement pre-employment, or post-conviction supervision of convicted offenders, is a set of tracings (charts) that can be systematically analyzed to make determinations of truthfulness or deception at rates that are greater than can be obtained by other methods. Other professions, such as medicine and education, use both diagnostic and screening methods in their respective fields. The scientific work that has been applied to these methods can also be applied to polygraph examination (Polygraph and lie detection, 2003). Among consumers of the information in both the medical and educational testing methods, there is a general implicit understanding that test results are helpful to professional decision making in that scientific test results have been shown to be significantly greater than chance, even if imperfect. This assumption is based on several predicate assumptions: that those administering and analyzing the tests have acquired advanced training and education; that these practitioners are qualified in their respective fields to select, administer, and interpret tests that will provide information that will assist the referring professionals to make better decisions.

Although signal detection theory[26] is not an integral part of this thesis, it is important to understand that the diagnostic analysis of polygraph tracings involves signal detection, particularly as an underpinning in the scientific work necessary to advance the field. Signal detection involves the diagnostician's being able to distinguish between signals and noise. McNicol called it "a theory about the way in which choices are made" (McNicol, 2005). Signal information is diagnostic information that the observer wants to see, and noise is any nonsignal information or background noise (Keating, 2005) that can make the identification of diagnostic information difficult. Clearly, extracting diagnostic information from the "raw" data of polygraph tracings involves the diagnostician—in this case a polygraphist or blind reviewer—making observations about two states and assigning an assessment of which state he observes. Test sensitivity (Polygraph and lie detection, 2003) involves the effectiveness with which signal information can be extracted and used to identify the issue of concern. Test specificity also involves the effectiveness with which the absence of signal information is determined and affects the ability of a test to determine when the issue of concern is not present. Harvey further describes this phenomenon in "Detection Sensitivity and Response Bias." He explains that the "detection performance" (diagnostics) is based on both a sensory process and a decision process. A simple yes or no can be the response as to whether or not a signal was present, or there can be a "rating of the confidence" that a signal was present. In the case of most polygraph TDA systems, the response is a yes or no, with the value of yes described in a positive or negative number. This involves a sensory process (sensitivity), as well as a decision process with a defined criteria parameter (in this case, the instructions contained on the hand-score sheet) (Harvey, 2003).

[25] Mr. Raymond Nelson assisted in the research question designs, as well as the computation of the tables and figures and the interpretation of the results and analysis.

[26] As one might deduce, signal detection theory had its early beginnings with those researching radar. Its psychological roots began in the 1950s and were primarily led by John A. Swets (Herbert, 2010). See Herbert for an insightful article about Mr. Swets and signal detection theory.

In signal detection theory, this sensory and response criteria process involves "hits" and "misses." That is, there is a hit when the diagnostician says yes to the signal that is present (hit rate), and a miss (false-alarm rate) occurs when the diagnostician says yes to a signal that is not present, meaning that noise was wrongly identified as a signal. Table 2 graphically displays this theory.

Table 2. Conditional Probabilities, Signal Detection Theory

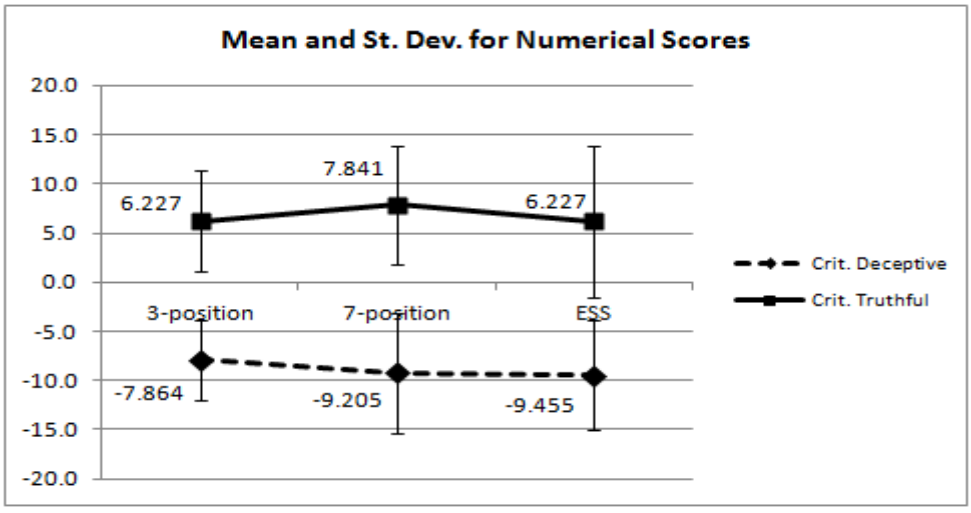
	"Yes"	"No"
Signal Present	Hit Rate (HR)	Miss Rate (MR)
Signal Absent	False Alarm Rate (FAR)	Correct Rejection Rate (CRR)

In polygraph, the FAR and MR are respectively known as false positive[27] and false negative.[28]

Analysis Method

Three-position and seven-position TDA numerical scores were transformed to ESS scores and subjected to a 2 x 3 ANOVA (criterion state x TDA system) for absolute magnitude of mean numerical scores. Transformation to a common numerical scale ensures that differences are not attributable to scale differences and are a reflection of differences in the effectiveness with which examiners extract diagnostic (signal) information using the three TDA systems.

Figure 1. Mean and Standard Deviations for Numerical Scores



[27] The false detection of something that is not actually present. In polygraph it is the incorrect decision that deception was practiced by the examinee (Krapohl,& Sturm, 2002).

[28] The failure to detect the presence of a particular event or item. A false negative in polygraph refers to the incorrect decision that deception was not practiced by the examinee (Krapohl & Sturm, 2002).

Table 3. 2 x 3 ANOVA Summary (Criterion State x TDA model) for Mean Scores

Source	SS[a]	df[b]	MS[c]	F[d]	P[e]	F crit .05[f]
TDA System	96.212	2	1.093	0.032	.969	3.031
Criterion state	284.379	1	2.154	0.063	.803	3.878
Interaction	44.576	1	44.576	1.294	.256	3.878
Error	8890.591	258	34.460			
Total	425.167	262				

[a]SS Sum of Squares

[b]df Degrees of Freedom

[c]MS Mean Square

[d]F The F Value

[e]p Probability Value

[f]F Critical Value of F with α= .05

The ANOVA analysis produced no significant differences—an indication that each of the TDA systems is capable of extracting similar signal (diagnostic) information from the raw data. That is, using any one of the three TDA systems, the polygraphist should be able to observe the criterion for truthfulness or deception, with no one system being more or less diagnostic.

Research Question #2

Results

Are there significant differences in criterion accuracy for the three-position, seven-position, and ESS TDA systems?

Criterion accuracy (validity) refers to how effectively the testing system places individual cases in the correct criterion category. In polygraph, the signals intended to be captured are the test results of deception indicated or no deception indicated. In the case of a single issue examination, such as a criminal investigation or event-specific incident, this measure (criterion) is the polygraphist's opinion about the examinee's deception or truthfulness corresponding to actual truthfulness (ground truth).

Within signal detection theory, one measure of stimulus is sensitivity, discussed below. Another measure within signal detection is response bias. This thesis does not research response bias, and it is left for future research; however, it is important to understand that the phenomenon exists. Response bias is the tendency of the diagnostician to choose one response over another. In other words, it is the tendency of a diagnostician to favor, that is, to be biased toward, the selection of one response over another. The more features available, the more opportunities for a diagnostician to become biased. Detection theory allows for determining or delimiting the distributions consistent with bias or sensitivity and specificity of a test measure. Sensitivity and bias taken together all lead to a decision system in which the stimulus classes reach equal- variance normal distributions for the decision variable, making them more meaningful. This decision system can then be tested using receiver operating characteristic curves, which then leads us graphically to the proportion of hits (signal) to the proportion of false alarms (noise). This becomes important in determining how to manipulate response bias—either through instruction or by use of a confidence rating (p value) (Macmillan, &

Creelman, 1996). More specifically, as response bias relates to polygraph scoring, the development of the ESS-TDA method is designed to reduce the response bias of polygraphists. Specifically, older TDA methods relied on more features and criteria to arrive at a final score. These attributes make the scoring methods difficult to learn (instruction) and more subjective (introducing response bias), with less interrater reliability (Blalock, Cushman, & Nelson, 2009). ESS utilizes the "bigger-is-better" rule, which means fewer features to score allows for ease in learning. Also, the ESS is the only hand-scoring method that has a p-value table. The use of the p-value addresses the second method of dealing with response bias—the use of confidence rating. Again, response bias is a topic for future research, but it is mentioned here to demonstrate that ESS addresses it and that the p-values associated with ESS allow for criterion selection that addresses levels of sensitivity.

Sensitivity is but one aspect of accuracy (validity). If deception is perfectly indicated whenever a lie is present, then the signal proves positive (deceptive) whenever a lie is present; the measure is positive for deceptive in all positive cases and no false negatives are produced; in other words, perfect sensitivity (*Polygraph and lie detection*, 2003).

Specificity is the other aspect of accuracy. If deception is absent, then the signal always shows negative and is therefore perfectly specific to deception; it produces no false positives. A test is more specific the greater the proportion of persons who appear nondeceptive on the test; in other words, perfect specificity (*Polygraph and lie detection*, 2003).

Analysis Method

The analysis method used was multivariate ANOVAs (criterion state x TDA system) for decisions with inconclusives (i.e., test sensitivity to deception and test specificity to truthfulness), inconclusive rates, and error rates.

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 .769}	.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}

Figure 2. Mean Plot for Decisions, Errors, and Inconclusive Results

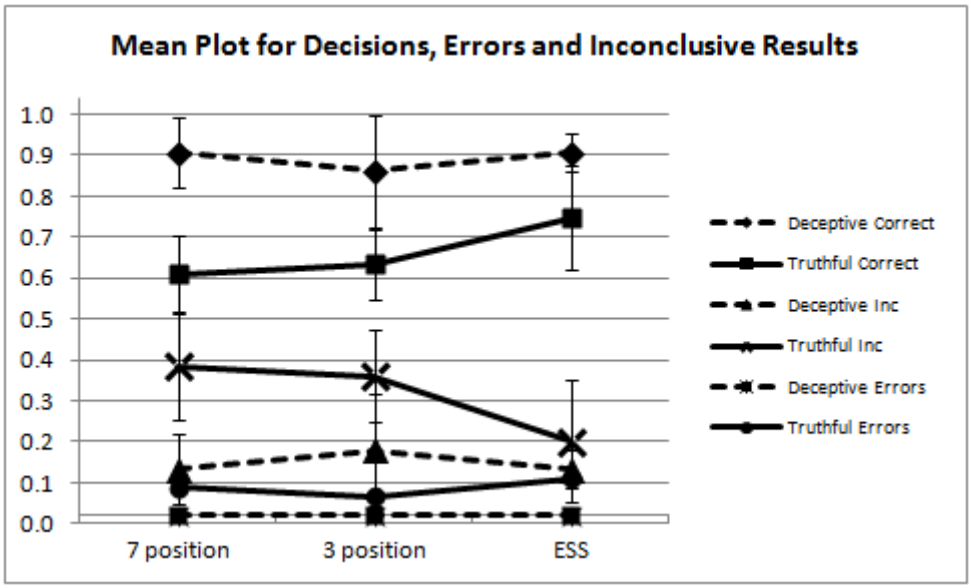


Table 5. Three-way (2 x 3 x 3) ANOVA Contrast for Test Accuracy (Criterion State x TDA System x Accuracy Dimension)

Source	S	df	MS	F	p	F crit.05
Criterion dimension	6.814	2	3.407	392.260	.000	3.168
Status	0.000	1	0.000	0.018	.893	4.020
TDA system	0.000	2	0.000	0.013	.987	3.168
Criterion dimension x status	0.504	2	0.252	28.985	.000	3.168
Status x TDA system	0.000	2	0.000	0.013	.987	3.168
Criterion dimension x TDA system	0.082	4	0.020	2.352	.065	2.543
Criterion dimension x status x TDA System	0.054	4	0.014	1.559	.198	2.543
Error	0.469	54	0.009			
Total	7.923	71				

The value of this three-way contrast is that it encompasses the entire experimental question; it provides greater degrees of freedom; and it provides more power than a series of two-way analyses.

There was no significance in this three-way interaction, which suggests no statistically significant differences in the accuracy of the three compared TDA systems. It is noted that the two-way interaction was significant for criterion dimension (x case status). This suggests that the different TDA systems may perform differently with criterion truthful and criterion deceptive cases.

In this instance, differences in criterion dimension are expected, in that it is hoped that error and inconclusive rates are lower than decision accuracy rates. This main effect did not undergo additional analysis. The most significant interaction in the three-way analysis was the two-way interaction of criterion dimension x case status. Again, this interaction supports the expectation that correct, inconclusive, and erroneous will not result in similar proportions. Because of this, two-way post-hoc ANOVAs were completed for each of the three dimensions of test accuracy: decisions, errors, and inconclusive results.

Figure 3. Mean Plot for Sensitivity and Specificity

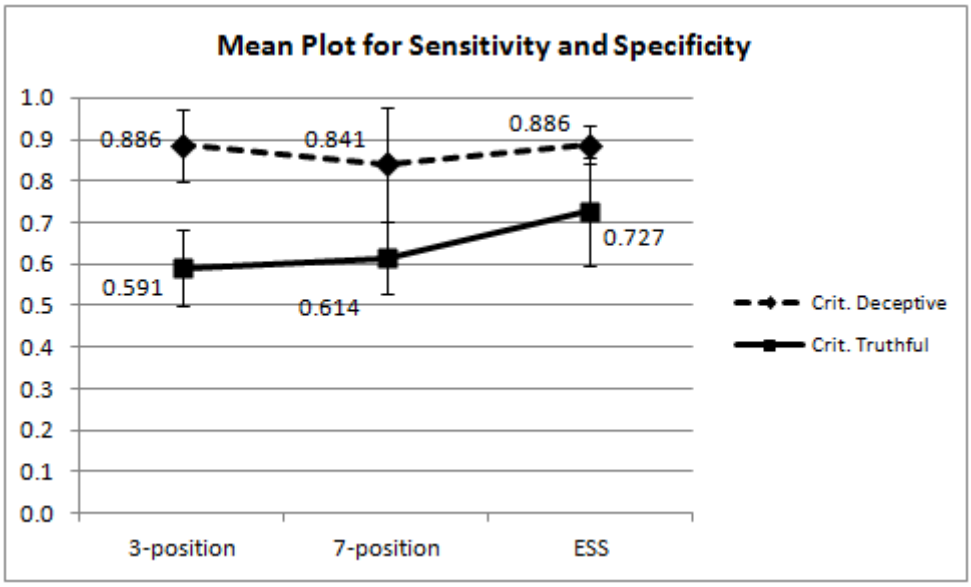


Table 6. Two-way ANOVA Summary (Case Status x TDA System) for Decision Accuracy, Including Inconclusive Results (i.e., Sensitivity and Specificity)

Source	SS	df	MS	F	p	F crit .05
TDA System	0.030	2	0.004	0.366	.698	3.555
Criterion state	0.310	1	0.026	2.557	.127	4.414
Interaction	0.019	1	0.019	1.841	.192	4.414
Error	0.182	18	0.010			
Total	0.358	22				

Neither the two-way interaction nor the main effects for case status or TDA system were significant for sensitivity and specificity.

Figure 4. Mean Plot for Inconclusive Results

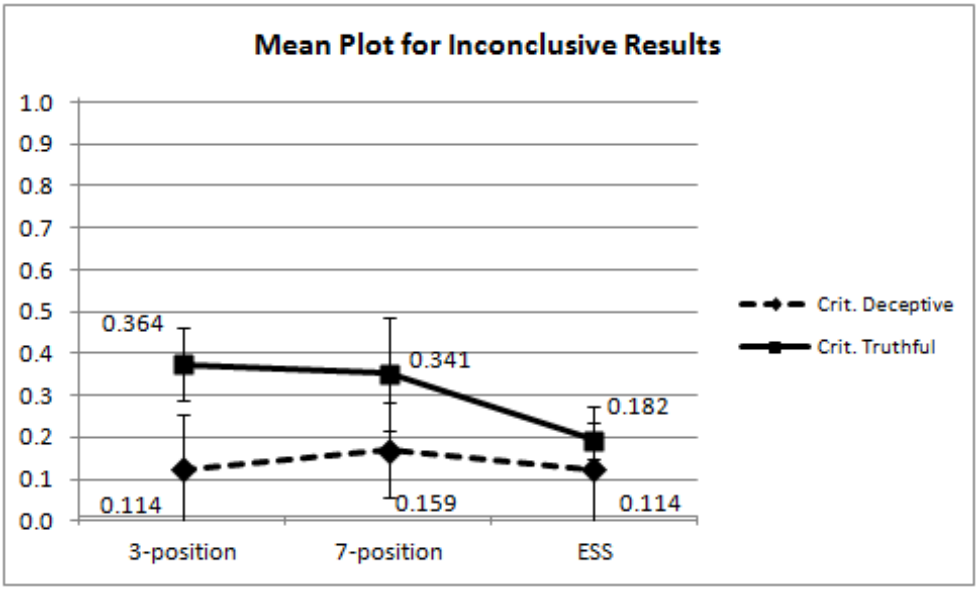


Table 7. Two-way ANOVA Summary (Case Status x TDA System) for Inconclusive Results

Source	SS	df	MS	F	p	F crit .05
TDA System	0.050	2	0.006	0.472	.631	3.555
Criterion state	0.167	1	0.014	1.043	.321	4.414
Interaction	0.034	1	0.034	2.534	.129	4.414
Error	0.240	18	0.013			
Total	0.251	22				

Neither the two-way interaction nor the main effects for case status or TDA system were significant for inclusive rates.

Figure 5. Mean Plot for Errors

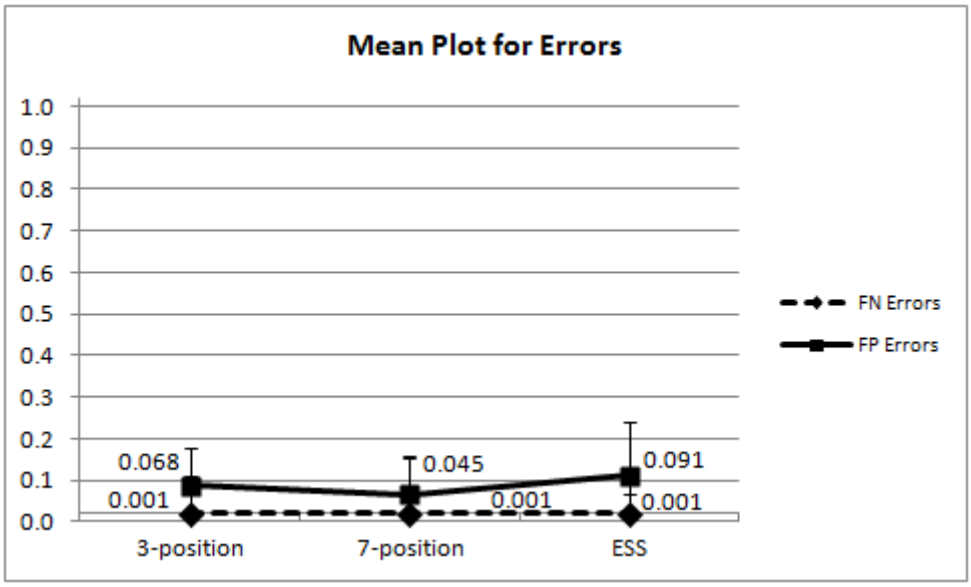


Table 8. Two-way ANOVA Summary (Case Status x TDA System) for Inconclusive Results

Source	SS	df	MS	F	p	F crit .05
TDA System	0.002	2	0.000	0.098	.907	3.555
Criterion state	0.027	1	0.002	0.855	.367	4.414
Interaction	0.002	1	0.002	0.782	.388	4.414
Error	0.048	18	0.003			
Total	0.031	22				

Neither the interaction nor the main effects of case status and TDA system were significant for errors.

The results of these analyses indicate that the three-position, seven-position, and ESS TDA systems produce different rates of correct, erroneous, and inconclusive results. However, there was no significance in the differences in the three TDA systems. It is noted that this may be a result of sample size and the size of the cohorts. Larger sample sizes and larger cohorts may produce significant differences.

No statistical power analysis was completed. Confidence intervals can be found in the table of means (Table 4).

It is noted that there is an absence of false-negative errors in this study. In a 2006 study, Krapohl reported a field study with a false-negative rate at 2.7% (Krapohl, 2006). The current error rate should be taken as statistically meaningless. It is unrealistic to expect this in field settings or larger studies. The result should be used with caution.

Research Question #3

Issue Posed

What is the effect on the accuracy of transforming three-position and seven-position scores to ESS scores?

Analysis Method

ESS scoring rules were applied to three-position and seven-position TDA systems, and a two-way ANOVA (TDA system x ESS transformation) was calculated.

Figure 6. Three-Position Score Transformed to ESS Scores

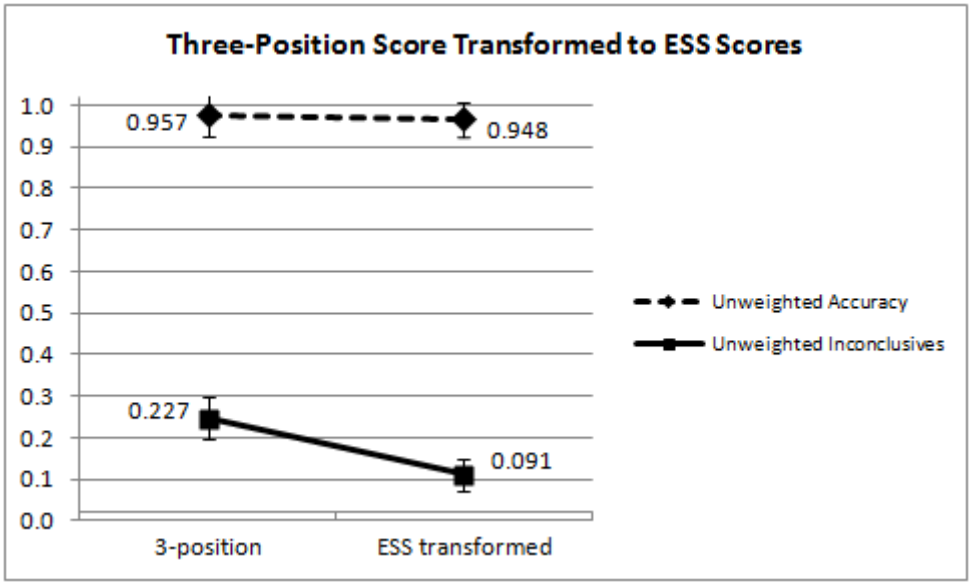


Figure 7. Seven-Positions Score Transformed to ESS Scores

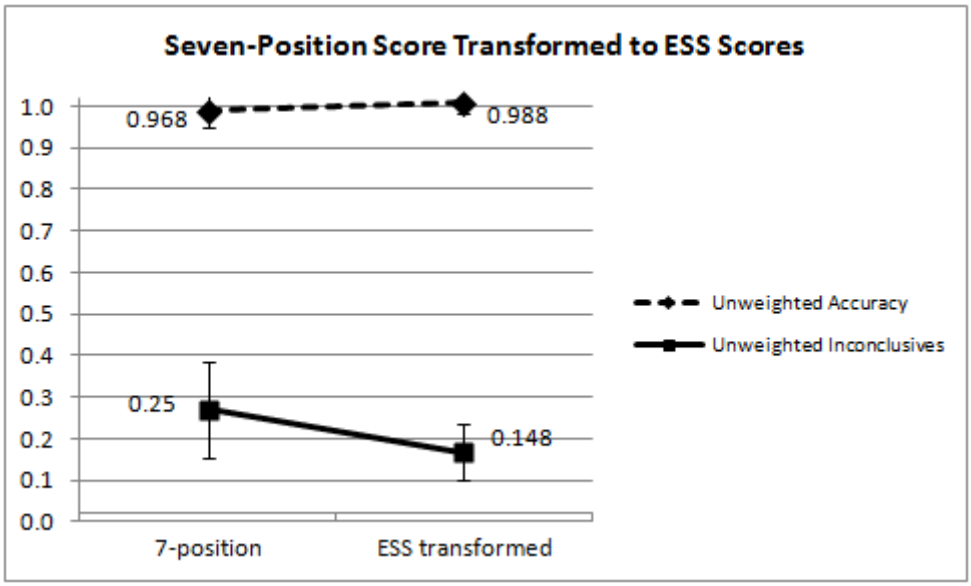


Table 9. Unweighted Accuracy

	Raw	ESS
3-position	.957 (.054) {.851 to >.999}	.948 (.041) {0.867 to >.999}
7-position	.968 (.037) {.895 to >.999}	0.988 (0.025) {.939 to >.999}

Table 10. Unweighted Inconclusives

	Raw	ESS transformed
3-position	.227 (0.052) {.124 to 0.33}	.091 (.037) {.018 to .164}
7-position	.250 (.114) {.026 to .474}	.148 (.068) {.014 to .281}

Table 11. Two-way ANOVA Summary (TDA System x ESS Transformation) for Accuracy

Source	SS	df	MS	F	p	F crit .05
Transformation	0.000	1	0.000	0.009	.928	4.747
TDA System	0.003	1	0.000	0.197	.665	4.747
Interaction	0.001	1	0.001	0.467	.507	4.747
Error	0.020	12	0.002			
Total	0.004	15				

No significant differences were found between the distributions of ESS scores and the transformed three-position and seven-position scores when a two-way ANOVA was conducted. Also, there were no significant differences in unweighted accuracy when transforming the scores of these TDA models to ESS scores.

Table 12. Two-way ANOVA Summary (TDA System x ESS Transformation) for Inconclusive Results

Source	SS	df	MS	F	p	F crit .05
Transformation	0.057	1	0.007	1.302	.276	4.747
TDA System	0.006	1	0.001	0.145	.710	4.747
Interaction	0.001	1	0.001	0.213	.653	4.747
Error	0.066	12	0.005			
Total	0.064	15				

There are also no significant differences between the three-position and seven- position TDA inclusive results when a two-way ANOVA was conducted for inconclusive results. A larger study may produce statistical power that could provide for expected improvement.

Research Question #4

How accurate are the combined 3-position, 7-position, and ESS TDA results? How accurate are the combined results when all scores are transformed to ESS scores? Is the difference significant?

Table 13. Accuracy of ESS, Three-position and Seven-position Scores Combined

	Raw scores	All scores transformed to ESS Scores
Unweighted Accuracy	.957 (.043) {.874 to >.999}	.961 (.040) {.883 to >.999}
Unweighted Inconclusives	.208 (.090) {.032 to .384}	.129 (.064) {.004 to .254}

Figure 8. Raw Scores and Transformed ESS Scores

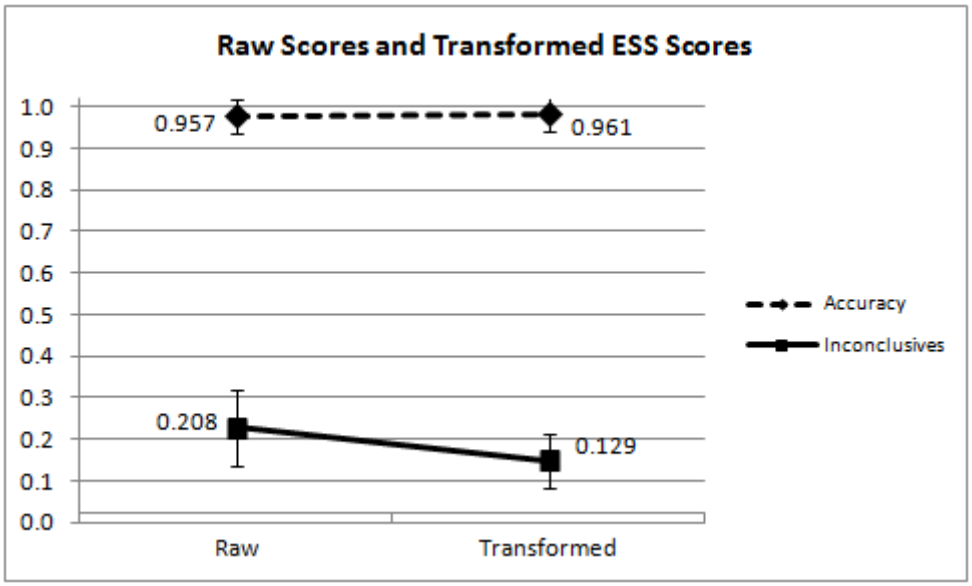


Table 14. Two-way ANOVA Contrast (Transformation x Accuracy Dimension) for Test Accuracy.

Source	SS	df	MS	F	p	F crit .05
Transformation	0.017	1	0.001	0.186	.669	4.062
Dimension	7.498	1	0.312	80.381	.000	4.062
Interaction	0.021	1	0.021	5.330	.026	4.062
Error	0.171	44	0.004			
Total	7.537	47				

Because it is known that the proportion of inconclusive differs from the proportion of correct, there is an expected significant main effect for accuracy dimension. ESS-transformed scores will produce different types of changes in decisions and inconclusive, as significant interaction for transformation and accuracy dimension suggests; decision accuracy increases and inconclusive results decrease.

One-way differences for decision accuracy were not significant [F(1,22) = 0.004, (p = 0.952)].

A larger sample may have found a significant difference in these results: one-way differences for inconclusive results were also not significant [F(1,22) = 0.522, (p = 0.478)].

Research Question # 5

Are there differences in accuracy that can be attributed to experience? Does more experience result in increased accuracy?

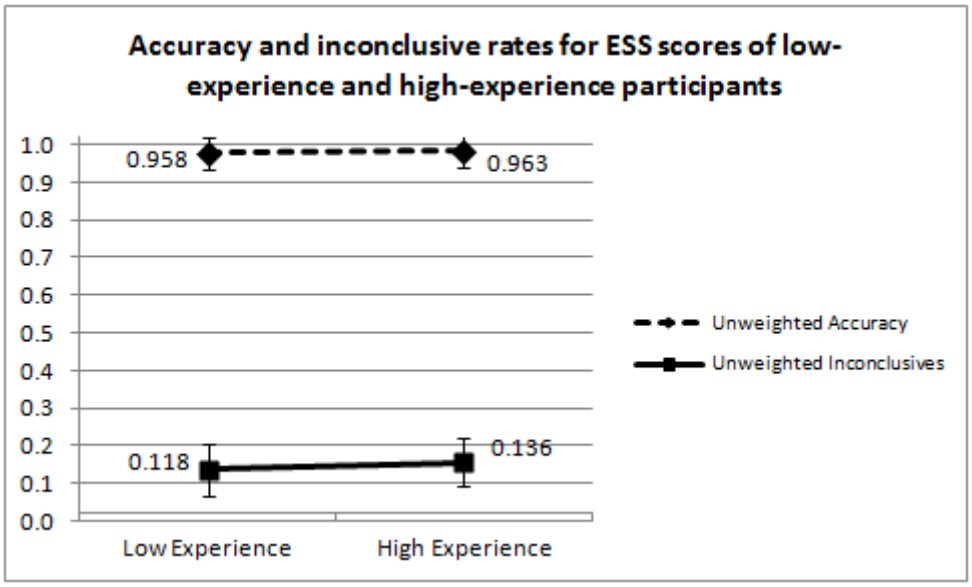
The average years of experience are 15. The standard deviation is three. The maximum years of experience are 33. The minimum years of experience are three. The median years of experience are 14, and the mode is 14. None of the participants is considered inexperienced.

For the purpose of this research, fewer than ten years is considered low experience and more than ten years is considered high experience.

Table 15. Accuracy and Inconclusive Rates for Low-Experience and High-Experience Participants

	Low Experience	High Experience
Unweighted Accuracy	.958 (.043) {.873 to 1.042}	.963 (.041) {.883 to 1.044}
Unweighted Inconclusives	.118 (.069) {<.001 to .253}	.136 (.064) {.010 to .262}

Figure 9. Accuracy and Inconclusive Rates for ESS Scores of Low-Experience and High-Experience Participants



Due to differences in sample size and an expected difference in decision and inconclusive rates, unbalanced one-way ANOVA were used.

Results between high- and low-experience participants were not significant for decision accuracy [$F(1,10) = 0.009$, ($p = 0.925$)]. Neither were results significant for inconclusive results [$F(1,10) = 0.037$, ($p = 0.851$)].

There was no effect for low or high experience in this sample data. That is, the low-experience participants scored polygraph charts using ESS with the same accuracy and inconclusive rates as high-experience participants. This outcome is consistent with that reported between inexperienced scorers and experienced scorers by Blalock, Cushman, and Nelson (2009) and Krapohl and Cushman (2006).

Research Question #6

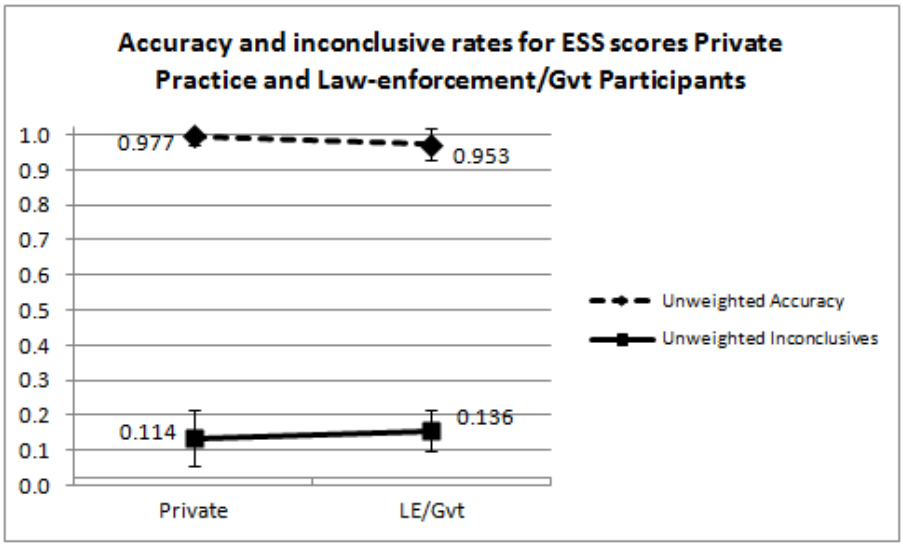
Does accuracy vary with the examiner's type of employment? Are there differences in accuracy between private examiners and those who work for law enforcement or government agencies?

One federal examiner was combined with the county/local law enforcement group for a combined group of government employees.

Table 16. Accuracy and Inconclusive Rates for ESS Scores of Private-Practice and Law Enforcement/Government Participants

	Private	LE/Gvt
Unweighted Accuracy	.977 (.026) {.926 to 1.029}	.953 (.045) {.865 to 1.04}
Unweighted Inconclusives	.114 (.079) {<.001 to .268}	.136 (.060) {.020 to .253}

Figure 10. Accuracy and Inconclusive Rates for ESS Scores of Private-Practice and Law Enforcement/Government Participants



Due to differences in sample size and an expected difference in decision and inconclusive rates, unbalanced one-way ANOVA were used.

Results were not significant for decision accuracy [F(1,9) = 0.228, (p = 0.644)], nor were results significant for inconclusive results [F(1,9) = 0.053, (p = 0.823)].

There was no effect for type of employment in this sample data, although a larger sample size may be expected to produce different results.

Discussion

Introduction

Polygraph has been used as a tool by the federal government, the military, and state and local governments for several decades. It has been and continues to be used as a successful instrument in national-security issues, homeland security, and the war on terror. Nevertheless, its detractors and those unfamiliar with its utility and successes, as well as disagreements and lack of foresight within the profession itself, have caused some

of the agencies and decision makers who do or could benefit from its use to be reluctant to rely on it. Some of this reluctance and even abandonment, in spite of the lack of alternatives, is due to outside pressure. The pressure can come in the form of political pressure, from uninformed law or policy, or from those who believe they have been wronged or harmed through the use of the polygraph. The challenge then has become multifold: Polygraph proponents must ensure that there is ongoing research to address the concerns and, in some instances, the valid arguments and criticisms of detractors (some of whom work in other scientific disciplines), and they must continue to move the profession onto a sound scientific foundation. Within the profession, infighting and lack of foresight and vision must be overcome for the sake of ensuring that this tool remains viable and valuable in its contribution to the homeland's safety and security, regardless of the form that proven instrumentation and technology takes. It is imperative that decision makers, policy makers, and other consumers who currently rely on the polygraph (as well as those who should) be educated by those within the profession who can and should undertake such a goal.[29]

Various studies and reviews have been undertaken in regard to hand-scored polygraph techniques. Two primary hand-scoring techniques in use today are the three-position and seven-position TDA systems. These two systems employ twelve scoring features for the purpose of assigning positive values (no deception indicated) or negative values (deception indicated) when the responses to relative questions are compared to the comparison questions. The rules (instructions) for assigning values are complex. The Empirical Scoring System uses observation of three scoring features for the purpose of assigning these negative and positive values. The instructions for assigning these values are simple and rely on the bigger-is-better rule.

The purpose of this study was to extend the research into the Empirical Scoring System to see whether it has additional value or is at least the equivalent of other hand-scoring techniques currently in use. Various research questions were posed, and through the use and analysis of raw data, comparisons were made between the three-position and seven-position scoring techniques, arguably two of the most highly utilized scoring techniques in the polygraph profession. These two techniques have been in use since the 1960s and are taught at the National Center for Credibility Assessment (NCCA),[30] as well as other private and government-funded polygraph schools across the United States and internationally.[31] Previous research on hand-scoring techniques was normative-based, while research on the ESS is empirically based and has allowed for the assignment of p-values to the technique. The intent of this study was to conduct additional research of the ESS, to further determine whether its design and method of use offer advantages over the compared techniques.

[29] Polygraph has a long history of infighting within the profession, as the author can attest to from his 25-plus years in the field. This infighting tends to revolve around scientific research and its importance to the trade. There are those within the field who believe that the profession need not be concerned about what detractors say about the validity and reliability of the polygraph. This side of the house tends to argue that "we know it works" and its utility is incontrovertible. The other side of the house argues that, for the field to survive, the polygraph must continue to build a strong scientific foundation. That is, in order to continue to serve its important role in national security and law enforcement, it must prove its validity and reliability so that its worth can be proven to policy makers and legislators in contrast to the naysayers' claims. This thesis falls on this side of the argument.

[30] NCCA is the Department of Defense's polygraph school, which all federal polygraphists attend as part of their initial training. It was formerly known as the Department of Defense Polygraph Institute (DODPI) and later as the Defense Academy for Credibility Assessment (DACA).

[31] Currently, the American Polygraph Association, the largest professional polygraph association in the world, accredits 16 schools in the United States and 13 international schools.

Discussion

The research design, primarily through use of ANOVA, was intended to measure several facets of the ESS as compared to the three-position and seven-position TDA systems. The study used 22 archival matched random samples of You-Phase examinations from the confirmed case archive at the Department of Defense. Eleven of these cases were confirmed truthful examinations. "Confirmed truthful" in the instance of these 11 examinations means that an alternative person was identified as a suspect or the examinee was exonerated, as there was evidence or a confession outside the opinion rendered by the specific polygraphist. Eleven matching confirmed deceptive examinations were also provided. "Confirmed deceptive" in the instance of these 11 examinations, means that there was evidence or a confession outside the opinion rendered by the specific polygraphist. As per the You-Phase protocol, which is part of the examination technique, these are single-issue examinations that contain two relevant questions and three comparison questions, as well as other procedural questions. The study participants were randomly selected and consisted of three groups (Cohorts 1, 2, and 3) of four scorers each. The first cohort utilized the ESS-TDA. The second cohort used the three-position TDA system. The third cohort used the seven-position TDA system. There were six research questions in the study.

The first question was to discover whether there were differences in the ability of each TDA system to extract diagnostic data from the provided examinations. This analysis was undertaken through use of ANOVA. The ANOVA analysis produced no significant differences between the three TDA systems; each was as capable as the others of extracting diagnostic information.

The next research question was whether there were differences in the three TDA systems for criterion accuracy (validity). This analysis was conducted through the use of multivariate ANOVAs and targeted inconclusives, inconclusive rates, and error rates. No significant differences were found in the three-way interaction. However, in the two-way interaction it was significant for criterion dimension. This suggests that the TDA systems may perform differently with criterion truthful and criterion deceptive cases. This interaction supports the expectation that correct, inconclusive, and error rates will not result in similar proportions. That is, the hit rate should be better than the miss rate and the indeterminate rate. The two-way interaction showed no significance for sensitivity or specificity, and this supports the expectation. Although the systems produced different rates of correct, errors, and inconclusive results, there are no significant differences in the three TDA systems. ESS seems to have a better specificity, lower inconclusive, and an equivalent error rate (which approached zero for all three TDA systems; again, an unrealistic result that should only be used with caution). It is noted that both the sample size and the size of the cohorts may have had an effect on this lack of significance. A larger sample size and larger cohorts may produce significant differences. What can be said as a result of this research is that ESS appears to have at least the same criterion accuracy as the three-position and seven-position TDA systems.

Transforming the three-position and seven-position TDA scores to ESS scores was conducted to determine whether there was an effect on the accuracy of the three- and seven-position scoring systems. This was accomplished through application of a two-way ANOVA. No significant differences were found between the distributions of the three TDA systems when the two-way ANOVA was conducted, which means that there is a high correlation between the three when transformed. There was no significant difference in unweighted accuracy as the result of transformation. In terms of inconclusive rates for the transformed three-position and seven-position scores transformed, no significant differences were found. This was an unexpected result. The expectations were that the inconclusive rates would be higher for both the three-position and seven-position TDA systems transformed to ESS. This unexpected result is likely the result of a small sample size. A larger study should produce statistical power that may provide the expected results.

A fourth research question looked at the accuracy of the result of the combination of the raw scores of all three TDA systems to ESS. There was a significant main effect for the accuracy dimension, and this was expected, given that it is known that the proportion of inconclusive will differ from the proportion of correct. Decision accuracy increased and inconclusive results decreased. One-way differences in decision accuracy were not significant; however, neither were one-way differences for inconclusive results. It is hypothesized that a larger sample may find significant differences.

The fifth question under study was to discover whether there were differences in accuracy based on level of experience. A one-way unbalanced ANOVA was utilized for analysis, and there were no significant differences for decision accuracy or inconclusive results based upon experience. Although there were no inexperienced participants in any of the cohorts, there were participants with low experience and participants with high experience. The results seem to show that there is no effect in the application of ESS scoring techniques based on years of experience in the field of polygraph scoring.

A final analysis was conducted to determine whether type of employment, private or government, had any effect on accuracy. An unbalanced one-way ANOVA found that there were no significant differences for decision accuracy or for inconclusive results. There was no effect for type of employment based on this research. However, it is again hypothesized that a larger sample size might produce different results. The analysis seems to indicate that type of employment has no effect on TDA diagnostics.

Limitations

Sample size was the primary limitation of the current study, both in terms of confirmed case sample, as well as the number of participants. Larger sample sizes would produce more statistical power, and it is hypothesized that they would have an impact on the significance of some of the findings of this study. Additionally, the study participants were experienced polygraphists who had attended a continuing education seminar and classroom instruction on ESS. It cannot be concluded that these cohorts are representative of the wider population of polygraphists. Another limitation is that it is not known how the confirmed cases came to be selected into the archive, other than being confirmed cases. The researchers intentionally used cases confirmed by extra-polygraph means, but one must consider that the selection may potentially lead to criterion accuracy rates that could be overestimated.

Recommendations for Future Research

Several recommendations for future study can be made as the result of this research. First, since the sample size was small, it is possible that the statistical power in a larger sample could reveal differences that escaped detection in this project. This larger sample size may be of interest in research by type of employment, decision accuracy, criterion accuracy, and other considerations. It is of note that government polygraphists, particularly federal government examiners, typically attend government-sponsored polygraph schools, while private examiners typically attend private schools, although many private examiners are retired government polygraphists. The results may reveal differences in instruction, expectations, types (quality) of exams conducted, or overall workload (number of tests conducted), among other possible variables that can then be studied.

Another aspect of test data analysis that may be of interest is the amount of time required to use the various types of hand-scoring techniques. These time studies can then be correlated to other demographic facets of the participants, again including age, experience, and type of employment. The p-value tables for ESS are well developed, although their significance to field polygraphists, decision makers, and other consumers is not well known.

Response bias is another issue that was not addressed by this study. It has a direct impact on sensitivity and should be further studied. Further research into the potential importance of this attribute of ESS

and its potential contribution to policy decisions should be undertaken. This research suggests that ESS can at least complement, if not supplant, the two compared TDA systems, and perhaps others, to increase the value of polygraph to homeland security and the war on terror. Further research should be conducted into this potential.

Conclusion

The first conclusion that can be drawn from this study is that ESS has at least as much diagnostic ability as the three-position TDA system and the seven-position TDA system, even taking into consideration the newness of the ESS to the polygraph profession. Despite the limitations of the small sample size, the study produces partial evidence and suggests that ESS has consistently high criterion accuracy. The study hints that ESS seems to have a better specificity, less inconclusive rates, and at least equivalent error rates as the three-position and seven-position TDA systems. Lastly, the study seems to support past research into ESS that the length of experience has no impact on the ability of the polygraphist to apply ESS scoring rules. This offers a particular advantage over the more complex scoring systems, which include more features and scoring rules, since graduates of polygraph schools seldom have time to ease into their new jobs. That is, the new graduate of a polygraph school can typically expect that soon after his assignment, he will undertake a polygraph examination that has high impact and consequences. The impact and consequences can literally save or cost lives, determine the future course of major tactical plans and actions, or forever change the lives of individuals.

It is imperative that those who can impact the use of the polygraph in the United States continue to pursue the lofty goal of sound and scientifically based lie detection techniques, procedures, instrumentation, and technology. Consideration must be given to programs and projects that will get the information about best practices into the hands of practitioners, decision makers, and consumers, some of whom have little knowledge about the abilities or contributions of the polygraph. There are decision makers—such as military officers, police chiefs, judges and prosecutors, government program directors, government officials and others—who make decisions based upon polygraph examinations and subsequent rendered opinions, who have never been educated about the polygraph. They do not know that there are methods available, based on well-founded chosen policy decisions, that will better provide them with the information that they want to have, taking into the account sensitivity, specificity, error rates, and inconclusive rates that ESS offers. It could well be worthwhile for the profession to develop educational seminars to inform stakeholders about these considerations. This could be a particularly worthwhile endeavor for the American Polygraph Association and the American Association of Police Polygraphists. Given that these two associations already have networks with various stakeholders, such as the Department of Defense, the International Association of Chiefs of Police, the National Sheriffs Association, and their state counterparts, short programs could be developed to introduce these ideas at their respective conferences and meetings and provide follow-up through articles in their widely circulated professional publications. It is also important to reach rank-and-file personnel who are actually in the field and may be unaware of the current state of the polygraph and best practices. The author is aware of state law enforcement academies that address the polygraph in basic courses as well as continuing education courses. There are analogous educational undertakings in the military and other programs through which these short, informational classes could be offered.

In terms of the profession itself, there must be a major internal push to continue the research that has been undertaken in the last several years. We must keep our eye on the target, and that target cannot be misidentified. As the National Research Council suggested in its seminal report, the concern must be on national security and, by implication, homeland security and the war on terror. If research into lie detection and other social sciences identifies better methods, instrumentation, technologies, and

techniques, then they must be further studied and embraced, if proven, even at the expense of letting go of what we know and what gives us comfort.

In terms of the present, careful consideration must be given as to how to keep current practitioners within the bounds of known best practices. Scientific and scholarly research and peer-reviewed articles are part of that equation; however, one must not lose sight of the polygraphist in the field whose primary concern is learning today what can be applied tomorrow. The science and research must be translated and presented in such a way that these individuals take an interest in it, understand it, and apply it.

Lastly, this study supports some of the findings of previous research into the Empirical Scoring System. It seems to support the position that ESS can complement the three-position TDA and the seven-position TDA systems and potentially others. This study did not find that ESS improved the scoring ability of the polygraphists. It does support the position that ESS offers the ability of polygraph consumers to choose their own tolerance for risk, something that is not readily available with other scoring systems.

This ability for the consumer to choose levels of risk when relying on the polygraph is important, but often not understood, and it can play a valuable role in homeland security and the war on terror.

> No theory is going to be inviolate. Let me put it clearly. The only kind of theory that can be proposed and ever will be proposed that absolutely will remain inviolate for decades, certainly centuries, is a theory that is not testable. If a theory is at all testable, it will not remain unchanged. It has to change. All theories are wrong. One does not ask about theories, can I show that they are wrong or can I show that they are right, but rather one asks, how much of the empirical realm can it handle and how must it be modified and changed as it matures? (Chadee, 2011, quoting Leon Festinger, 1987)

References

Alder, K. (2007). *The lie detectors: The history of an American obsession*. New York: Free Press.

Ball & Gillespie Polygraph. (n.d.). William Marston. Retrieved January 16, 2012, from http://www.lie2me.net/thepolygraphmuseum/id17.html

Blalock, B., Cushman, B., & Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. *Polygraph*, 38(4), 281–88.

Chadee, D. (2011). *Theories in social psychology*. Malden, MA: Wiley-Blackwell. C.I.A.. (n.d.). *DHRA Web site | Home*. Retrieved October 11, 2011, from http://www.dhra.mil/perserec/espionagecases/cia.htm

Collins, C. (2004). OSI provides the truth in Baghdad. *Learn How to Pass (or Beat) a Polygraph Test | AntiPolygraph.org*. Retrieved January 17, 2012, from http://www.antipolygraph.org/articles/article-044.shtml

DACA. (2006). Test Data Analysis. *Psychological detection of deception analysis II— Course #503*. Lecture conducted from Defense Academy for Credibility Assessment, Ft. Jackson, South Carolina.

Department of Defense Polygraph Program: FY 2002 Report to Congress. (2003). *Federation of American Scientists*. Retrieved January 15, 2012, from http://www.fas.org//sgp/othergov/polygraph/dod-2002.html

Earley, P. (1997). *Confessions of a spy: The real story of Aldrich Ames*. New York: G.P. Putnam's Sons.

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

Harvey, L. (2003). Detection sensitivity and response bias. Retrieved December 17, 2011, from psych.colorado.edu/~lharvey/p4165/p4165_2003_spring/2003_Spring_pdf/P4165_SDT.pdf

Herbert, W. (2010). John A. Swets: A signal idea, a singular life. *Observer* 23(7), 150–165. Retrieved January 15, 2012, from http://www.psychologicalscience.org/index.php/publications/observer/2010/september-10/john-a-swets-a-signal-idea-a-singular-life.html

History. (n.d.). *Joe Klingensmith—Polygraph service*. Retrieved September 12, 2011, from http://polygraphservice.us/History.htm

Hoffman, D., & Stober, I. (2001). *A convenient spy: Wen Ho Lee and the politics of nuclear espionage*. New York: Simon & Schuster.

Keating, P. (2005). D-prime (signal detection) analysis. *UCLA Linguistics*. Retrieved January 17, 2012, from www.linguistics.ucla.edu/faciliti/facilities/statistics/dprime.htm

Kleiner, M. (2002). *Handbook of polygraph testing*. San Diego, CA: Academic Press.

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

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

Krapohl, D., & Dollins, A. (2003). Relative efficacy of the Utah, Backster, and federal scoring rules: A preliminary investigation. *Polygraph* 32(3), 150–65.

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

Krapohl, D., Stern, B.A., & Bronkema, Y. (2009). Numerical evaluations and wise decisions. *Polygraph,* 38(1), 57–69.

Krapohl, D., & Sturm, S. (2002). Terminology reference for the science of psychophysiological detection of deception. *Polygraph,* 31(3), 154–239.

Lerner, B., & Lerner, E. (2006). Polygraph, case histories—eNotes.com. *eNotes - Literature Study Guides, Lesson Plans, and More*. Retrieved October 10, 2011, from http://www.enotes.com/polygraph-case-histories-reference/polygraph-case- histories

Macmillan, N., & Creelman, C. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review,* 3(2), 164–70.

McNicol, D. (2005). *A primer of signal detection theory*. Mahwah, N.J.: L. Erlbaum Associates.

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., et al. (2012). Using the Empirical Scoring System. *Police Polygraphist Digest,* 4, 9–15.

Pentagon's intelligence arm steps up lie detecting efforts on employees. (2008, August 24). Retrieved February 22, 2012, from http://www.foxnews.com/story/0,2933,409502,00.html

Pincus, W. (1994, August 10). Ames says more polygraphs might have deterred him. *Washington Post*, p. 8.

The Polygraph and Lie Detection. (2003). Washington, D.C.: National Academies Press. Raskin, D. C., Barland, G., & Podlesny, J. A. (1978). *Validity and reliability of detection of deception*. Washington, D.C.: National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Dept. of Justice.

Scientific validity of polygraph testing: A research review and evaluation. (1983). Washington, D.C.: Congress of the U.S., Office of Technology Assessment.

Senter, S. (2008). Polygraph research: Value, shortcomings and perspectives. *Polygraph*, 37(4), 277–80.

Shelby, R. (2001). Intelligence and espionage in the 21st century. *Conservative Policy Research and Analysis*. Retrieved October 11, 2011, from http://www.heritage.org/research/reports/ 2001/05/intelligence-and-espionage-in- the-21st-century

Trovillo, Paul. (1972). The history of lie detection (part 1). *Polygraph,* 1(2) 46–75.

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

Weaver, R. (1980). The numerical evaluation of polygraph charts: Evolution and comparison of three major systems. *Polygraph,* 9(2), 94–108.

Weinberger, S. (2008). Handheld lie detector goes to war. *Wired.com*. Retrieved January 16, 2012, from http://www.wired.com/dangerroom/2008/04/handheld-lie-de/

Weiner, T. (1994, August 10). Polygraph detected lies by Ames, F.B.I. finds. *New York Times*, p. 6.

Wen Ho Lee Case Study. (2008). Retrieved August 22, 2011, from www.pherson.org/PDFFiles/WHLCaseStudy.pdf

Wen Ho Lee's Problematic polygraph. (2000, February 4). *CBS News*. Retrieved August 22, 2011, from http://www.cbsnews.com/stories/2000/02/04/national/main157220.shtml