Polygraph

VOLUME 37 2008 NUMBER 3

Contents

Published Quarterly

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

Air Force Modified General Question Test Validation Study

Stuart Senter, James Waller and Donald Krapohl

Abstract

This study evaluated the effectiveness of the Air Force Modified General Question Test, a polygraph format commonly used in the Federal Government. A mock crime scenario was used that required deceptive participants to place a simulated bomb next to a road. Decision accuracy using the Air Force Modified General Question Technique was significantly above chance levels (defined as .500) for total decisions (.838) and excluding no opinion decisions (.849). In addition, decision accuracy was significantly above chance levels for both truthful (.917) and deceptive (.758) participants. A total of one no opinion decision (.015) was produced. This study provides evidence for the effectiveness of the Air Force Modified General Question Test format in terms of identifying truthful and deceptive participants. Further research should be conducted to explore the effectiveness of different variants of this technique, including different scenarios, question types, and question numbers.

Introduction

The Air Force Modified General Question Test (AFMGQT) is a polygraph technique that is widely used across the United States Federal government for a variety of purposes. The AFMGQT is used in both criminal specific and counterintelligence screening purposes. A key principle employed by the AFMGQT is the placement of a comparison question as the first evaluated question in the series. Cullen and Bradley (2004) demonstrated that this was a critical factor in the diagnostic value of the polygraph test. Despite its prominent use and integration of defensible principles, no published research exists on the validity the AFMGQT, an empirical void the present study sought to fill.

For the purposes of this study, accuracy was defined in two ways: total accuracy and definitive accuracy. Total accuracy was considered in the context of all possible instances. In credibility assessment, there are commonly three types of decisions that can be rendered regarding the truthfulness of an individual; these decisions are truthful, deceptive, and no opinion (not enough definitive information to make a decision of truthful or deceptive). The calculation of total accuracy divided the number of correct decisions by the number of correct, incorrect, and no opinion decisions, per the following formula:

$$\text{Total Accuracy} = \frac{\text{correct decisions}}{\text{correct decisions + incorrect decisions + no opinion decisions}}$$

Definitive accuracy only integrated those instances where a decision of truthful or deceptive has been rendered, excluding no opinion decisions. Thus, definitive accuracy was calculated using the following formula:

Acknowledgments

The views expressed in this report are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

$$\text{Definitive Accuracy} = \frac{\text{correct decisions}}{\text{correct decisions + incorrect decisions}}$$

These accuracy calculations were collapsed across performance in the truthful and deceptive experimental conditions, and were also calculated for the two groups individually.

Based on binomial power calculations, it was determined that 28 participants per cell would produce a power of .85 to detect a proportion of correct decisions of .75 at the .05 level of significance (UCLA Department of Statistics, 2004), based on an effect size of .25 (Cohen, 1988, p. 147). The effect size calculations were based on the .75 proposed proportion of correct decisions versus a chance level of .50. Using Cohen's approach, the effect size was simply .75-.50 = .25. The .75 proportion of correct decisions was based on specific-issue polygraph performance, from multiple laboratory and field-based studies (National Academy of Sciences, 2003). Thus, a minimum 60 participants would be used in the present study, including 30 truthful and 30 deceptive participants.

Method

Participants

Only participants who had not previously taken a polygraph examination were allowed to participate in the study. Participants were 69 United States Army basic trainees at Fort Jackson, SC. Thirty-six of these participants were assigned to the truthful condition and thirty-three were assigned to the deceptive condition. Twenty-nine of these participants were female and forty were male. Ages for these participants ranged from 17 to 38, with an average age of 20.8 (SD = 4.5).

Polygraph Examiners

Twelve polygraph examiners conducted polygraph examinations for this study. Polygraph examiners were employed by the federal government. Each examiner possessed a minimum of six years of operational polygraph experience and at least one year of polygraph instruction experience. With the exception of four cases, there was one blind scorer who made the final decision for each polygraph examination. This blind scorer had over 25 years of federal polygraph experience. In the other four cases, two polygraph examiners with over 10 years of federal polygraph experience blind scored and produced the final decisions for two cases each.

Design

The experiment took place over four days. Twenty participants were available each day for the study. Pairs of participants were randomly assigned to the truthful and deceptive condition using a block randomization scheme. Participants completed the experiment in pairs due to the "Battle Buddy" concept that is mandated for basic trainees on Fort Jackson (Donald J. Krapohl, personal communication, March 1, 2006). In essence, this policy requires basic trainees to maneuver around the base in pairs. For each day, five of the ten pairs were randomly assigned to the deceptive condition, with the other five pairs assigned to the truthful condition.

Apparatus

Lafayette (Lafayette, IN) and Axciton (Houston, TX) computerized polygraphs were used, with one examiner using an ink-based analog polygraph. All polygraph instrument use was based on personal preference of each examiner. Computerized instruments were connected to desktop or laptop computers and operate within a Microsoft Windows interface. Each polygraph included two corrugated rubber tubes for monitoring thoracic and abdominal respiration, a standard blood pressure cuff for monitoring cardiovascular information, and two disposable Ag/AgCl sensors with conductance gel for monitoring electrodermal activity from the hands.

Procedure

The experimental procedure was modeled after one that has been used with great success for many years at the University of Utah (Kircher & Raskin, 1988). Much of the success of the procedure is attributed to the complexity and level of engagement that is required of participants. The study was conducted at a federal training facility on Fort Jackson, SC. Participants were initially seated in a large room where they were allowed to watch television or read while they waited for the opportunity to participate in the experiment. Participant pairs were called by name and asked to step outside of the room. Participants were then told to go through a door that led to a long hallway. They were directed to go to a door with a white envelope taped to it, and that the envelope would contain their instructions. These instructions directed participants (if they chose to participate in the study) to first read and fill out an informed consent form, and if they agreed to participate in the study after reading the informed consent, to press play on the tape player (Note that the polygraph was referred to as a 'lie detector' in the participant instructions. Although the term is technically inaccurate, 'lie detector' was used in the participant instructions because it is a term generally understood by the lay public).

The room that the pairs entered contained a chair and a desk with a cassette tape player, two informed consent forms, with a pen placed upon each of them. Due to logistical constraints, the rooms also contained a polygraph, polygraph components, a polygraph chair, and a desktop computer. Two such rooms were used in this capacity, with a monitoring room in between from which experimental staff could monitor and record their progress. Pairs received their instructions together in one of the two rooms.

After completing the informed consent forms, participants pressed 'PLAY' on the tape player to hear their instructions. For deceptive participants, the tape-recorded instructions directed participants to obtain a simulated bomb from a hidden location outside of the building and to plant the device at a nearby location across the street that is commonly used for training purposes. They were instructed to place the device in a box next to a dirt road that is in this training area. Participants were instructed to arm the bomb and leave the area. When the participants moved to a safe distance (approximately 30 yards) from the site, an M116A1 Hand Grenade Simulator was detonated by an U.S. Army ordnance specialist. This ordnance produces a 125 decibel blast at a distance of 75 feet (Miller & Ruppert, 2004). The purpose of the detonation was to lead participants to believe the bomb they placed may have exploded. The hand grenade simulator was dropped by the ordnance specialist and detonated in a 2-foot hole surrounded by three levels of sand bags, in a hidden location approximately 10 yards from the location where the participants planted the simulated bomb. The grenade pit was located across the dirt road by which the participants were to place the simulated bomb. The tape-recorded instructions clearly instructed participants not to cross this dirt road, to prevent participants from coming close to the grenade pit or from seeing the ordnance specialist. No participants violated this instruction. In addition, participants were under constant (covert) observation while outside of the training facility, ultimately to ensure their safety, and also to verify that participants followed instructions.

The tape-recorded instructions told participants they had 15 minutes to find and plant the simulated bomb and that they were to return to the room where they received their instructions when they had completed this task. Also, the instructions told participants to prepare an alibi or excuse for being outside, should they be stopped or questioned by anyone. The instructions imparted that someone would meet them and escort them to an area where they would take a lie detector test, focusing on the bomb placement. The taped instructions relayed that if questioned by anyone the participant should indicate that they know that a bomb was placed and detonated nearby, but that they know nothing beyond that. Deceptive participants were instructed to appear truthful during the lie detector test.

Truthful participants were instructed, also via tape-recorded instructions, to leave the contact room and go to a gazebo that serves as a break area behind the training facility. Participants were told to remain at the break area for approximately 10 minutes and to return to the contact room in 15 minutes where they were escorted to a lie detector test (the extra five minutes provided time for participants to travel to and from the break area). The recorded instructions informed truthful participants that a bomb was placed and detonated in a nearby location, but no other details were provided to them. Truthful participants were told to cooperate with the lie-detector operator and to be truthful during the testing process.

Relative to the detonations, the explosions were audible within the building, though not discernable from other training-related explosions that periodically took place near the facility. No truthful participants reported hearing the explosion during the debriefing.

All participants were then individually escorted to a polygraph suite, where they were given a polygraph examination. Prior to data collection, a pretest interview lasting approximately 45-60 minutes, was carried out by a polygraph examiner. The polygraph pretest process involved a structured interview covering the following areas: an overview of the polygraph process, administration of a brief medical/biographical questionnaire (including gathering of personal history), a brief introduction to the polygraph instrument, its allied components (e.g., corrugated rubber tubes, etc.), and the physiological responses produced when someone lies, a brief discussion of the case facts and a review of the questions to be presented.

The polygraph question list included four types of questions; irrelevant questions, sacrifice relevant questions, relevant questions, and comparison questions. Irrelevant questions are non-emotion evoking questions that were used as buffer items at the beginning of the question sequence. The sacrifice relevant question was also placed toward the beginning of the question sequence and asked whether the participant intended to be truthful about their involvement in the bomb placement. Irrelevant and sacrifice relevant questions were not used for scoring purposes. Relevant questions related to whether the individual placed or participated in the bomb placement. Comparison questions related to previous instances of lying in different contexts. These two question types were used in the polygraph decision-making process described below.

Participants were told, via the tape-recorded instructions, that if the lie detector results indicated that they had been truthful, they would be allowed to complete the process without consequence. Participants were also told that, if found deceptive by the lie detector, they would have to stand before their drill sergeant, their unit, and the staff of the training facility and give a speech on honesty, integrity, and loyalty, tying in the mock crime that they had completed. This punishment was not actually administered to participants. This public speaking element is a common form of punishment applied by drill sergeants to troops found guilty of wrongdoing (Harold L. Palmer, personal communication, April 6, 2004). Fear of public speaking is also a fairly widespread form of anxiety, and represents an area that has been thoroughly explored in the behavioral literature (Addison, Clay, & Xie, 2003; Anderson, Rothbaum, & Hodges, 2003; Harb, Eng, & Zaider, 2003; Savitsky & Gilovich, 2003; Zohar, Livne, & Fine, 2003; many others). The purpose of this hypothetical punishment was to instill a strong sense of jeopardy in connection to the test outcome, in an effort to model the stakes inherent in real world testing.

Prior to the actual polygraph data collection process, participants were seated in a Lafayette adjustable arm polygraph chair (item number 76871), and sensors were placed on them. Two corrugated rubber tubes were attached to the participant's chest and abdominal areas. The blood pressure cuff was placed on the participant's bicep over the brachial artery. The Ag/AgCl sensors were attached to opposite sides of the palm of the participant's hand.

Following sensor placement, participants were asked to write a number from four to seven (inclusive) on a piece of paper. The participant was then asked about what number they wrote on the paper (e.g., regarding the number you wrote, was it the number 3?), and to lie (by saying 'No') when asked about the number they chose. Questions began with the number that was three less than the number the participant selected (e.g., one if the number four was selected), and continued in ascending order until the number that was two greater than the selected number was reached (e.g., six if four was selected). These questions were asked every 20-25 seconds, with the total process requiring approximately three minutes. This process follows standard field practice, and allows the examinee to become accustomed to the testing process.

After this preliminary test, the questions listed in Table 1 were presented to the participant, with the polygraph components attached. Questions were presented every 20-25 seconds, requiring approximately three minutes of data collection time. After the presentation of each question list, the examinee was provided a break of a few minutes during which the blood pressure cuff was deflated. The question list was presented three times. The data collection process required approximately 15-20 minutes.

Table 1. Polygraph Test Questions.

I1 Are the lights on in this room?

I2 Are you now sitting down?

SR Regarding the placement of that bomb, do you intend to answer truthfully each question about that?

C1 Did you ever make a promise that you had no intention of keeping?

R1 Did you participate in placing that bomb near that road?

C2 Did you ever tell a lie to someone who trusted you?

R2 Did you place that bomb near that road?

C3 Did you ever lie to take advantage of a friendship?

Alternative comparison questions:

Did you ever blame someone for something you did?

Did you ever cheat anyone out of anything?

After the data collection process, the polygraph charts were printed out and then evaluated by the polygraph examiner. Following field practice, the charts were then provided to a blind scorer who also evaluated the charts and decided whether the participant was truthful or deceptive. The blind scorer produced the final decision for each examinee, following recommendations provided by Iacono (1991). Generally, sympathetic physiological responses (e.g., respiratory suppression, electrodermal amplitude, and cardiovascular amplitude) to relevant and comparison adjacent question pairs were compared, within the respiratory, electrodermal, and cardiovascular channels. Because each relevant question was adjacent to two comparison questions, the comparison question producing the larger sympathetic reaction for each channel was compared to the relevant question response. Larger sympathetic responses to the relevant question in a pair resulted in the assignment of a negative value (e.g., -1, -2, or -3, depending on the magnitude of the difference). Larger sympathetic responses to the comparison question in a pair resulted in the assignment of a positive value (e.g., +1, +2, or +3). No measurable differences between the response magnitudes of the two questions resulted in the assignment of a 0. The scores assigned to each of the two relevant-comparison question pairs were to be summed across all presentations. A decision of deceptive was produced if the total for either question pairing was -3 or lower. A decision of truthful required values of +3 or higher for both question pairs. A no opinion decision was rendered in all other cases. If a no opinion decision was produced, the polygraph examiner collected three additional charts using the same question list. In such cases, the decision process was repeated by the blind scorer who produced the final decision.

Following the polygraph process, participants were then fully debriefed by an experimenter. All participants were thanked for their participation and were provided more information regarding the importance of their participation in the project. Deceptive participants were assured that they had, in no way, committed a crime or an act of terrorism. They were told that their actions were crucial toward the evaluation of a new credibility assessment technology, and that they should be proud of their contributions to the research effort. Finally, all participants were asked not to disclose any details of their participation for at least a year, to avoid any contamination of subsequent participants in the continuing series of research studies.

Data analysis focused on the accuracy of the polygraph decision as compared to a known ground truth. Decision accuracy was assessed using total accuracy and definitive accuracy, as described earlier, based on both human scoring. The mock crime scenario was considered to be validated and suitable for PCASS evaluation if the total accuracy rate met or exceeded .700, and if the definitive accuracy rate met or exceeded .800. These numbers were selected based on previous polygraph research compilations (National Academy of Sciences, 2003). Statistical significance (compared to chance) was assessed using proportion tests (Bruning & Kintz, 1997).

Results

A total of 76 participants began the study, with 38 of these assigned to the truthful condition and 38 assigned to the deceptive condition. Four deceptive participants were eliminated from the study due to experimenter error. Two of these occurred because the simulated bomb was not placed in the proper location and the participants were unable to locate it. Two were eliminated because the instruction tape was not rewound and the participants were confused by the instructions. One deceptive participant confessed to completing the mock crime during the pretest process and was eliminated from the study. Two truthful participants were eliminated due to the project as a whole running out of time late in the day. Thus, 36 truthful participants and 33 deceptive participants successfully completed the study, for a total of 69 participants.

The proportion of agreement for decisions produced between original examiners and blind scorers was .93. The correlation between the two groups of decision makers was *r* = .94. Calculation of Kappa, a statistic used to measure inter-scorer agreement (Viera & Garrett, 2005), given the possibility of chance agreement, resulted in a value of .75. The proportions of agreement, correlation coefficient, and Kappa for these pairwise comparisons were significantly above chance levels (all *ps* < .05).

Tables 2 and 3 show the decision accuracy results, by total and definitive accuracy, respectively. Effect size calculations using Cohen's (1988) approach for total and definitive accuracy are shown in Table 4. The proportion of correct decisions for truthful (z = 4.0, p < .001), deceptive (z = 3.0, p < .01), and the collective total (z = 4.9, p < .0001) were significantly above chance levels (50% or .50) for the original examiner. The three categories also significantly exceeded chance levels for the blind scorer (z = 5.0., p < .0001, z = 3.0, p < .01, and z = 5.7, p < .0001, respectively). For original and blind scorer decisions, decision accuracy for truthful and deceptive participants did not differ significantly (all *ps* > .05). In addition, decision accuracy between original examiners and blind scorers did not differ significantly for truthful, deceptive, or total comparisons (all *ps* > .05).

Table 2. Total Accuracy Rates for Original Examiners and Blind Scorers as a Function of Participant Veracity.

	Truthful			Deceptive			Total		
Decision Method	Cor	Err	NO	Cor	Err	NO	Cor	Err	NO
Frequency									
Original Examiner	30	3	3	25	6	2	55	9	5
Blind Scorer	33	3	0	25	7	1	58	10	1
Proportion									
Original Examiner	.833	.083	.083	.758	.182	.061	.797	.130	.073
Blind Scorer	.917	.083	.000	.758	.212	.030	.841	.145	.015
Average (unweighted) Decision Accuracy									
Original Examiner							.796	.133	.072
Blind Scorer							.838	.148	.015

Note: Cor = correct decision, Err = erroneous decision, NO = no opinion

Table 3. Definitive Accuracy Rates for Original Examiners and Blind Scorers as a Function of Participant Veracity.

Decision Method	Truthful	Deceptive	Total
Original Examiner	.909	.807	.859
Blind Scorer	.917	.781	.853
Average (unweighted) Decision Accuracy			
Original Examiner			.858
Blind Scorer			.849

Note: Definitive Accuracy excludes no opinion decisions from accuracy calculations. Cor = correct decision, Err = erroneous decision, NO = no opinion

Table 4. Effect Sizes for Total and Definitive Accuracy Rates as a Function of Participant Veracity for Blind Scorer Decisions.

Accuracy Type	Truthful	Deceptive	Total
Total	.417	.258	.341
Definitive	.417	.281	.353

The same pattern of results held with the original examiner for definitive accuracy, with performance for truthful ($z = 3.4$, $p < .001$), deceptive ($z = 4.7$, $p < .0001$), and the collective total ($z = 5.8$, $p < .0001$) significantly exceeding chance levels. This was also the case with the blind scorer ($z = 3.4$, $p < .01$, $z = 4.0$, $p < .01$, and $z = 5.3$, $p < .001$, respectively). As with total accuracy, no differences were found between truthful and deceptive performance or between original examiners and blind scorers (all *ps* > .05).

Table 5 displays the number of examinations and individual accuracy rate produced by each polygraph examiner. Pairwise comparisons showed that the difference in total accuracy was significant between Examiner 1 and Examiner 7, ($z = 3.2$, $p < .01$). However, given that Examiner 7 conducted only a single examination, it is difficult to make assertions from this finding.

Table 5. Total Accuracy Rates Produced by Polygraph Examiners using Blind Scorer Results.

	Truthful			Deceptive			Total		
Examiner	Cor	Err	NO	Cor	Err	NO	Cor	Err	NO
Frequency									
1	6	0	0	3	0	0	9	0	0
2	0	1	0	4	0	0	4	1	0
3	5	1	0	1	1	0	6	2	0
4	4	0	0	0	0	0	4	0	0
5	3	0	0	3	1	1	6	1	1
6	2	0	0	2	0	0	4	0	0
7	0	0	0	0	1	0	0	1	0
8	2	0	0	1	0	0	3	0	0
9	3	0	0	1	0	0	4	0	0
10	2	0	0	4	2	0	6	2	0
11	1	1	0	4	1	0	5	2	0
12	5	1	0	2	0	0	7	1	0
Proportion									
1	1.000	.000	.000	1.000	.000	.000	1.000	.000	.000
2	.000	1.000	.000	1.000	.000	.000	.800	.200	.000
3	.833	.167	.000	.500	.500	.000	.750	.250	.000
4	1.000	.000	.000	.000	.000	.000	1.000	.000	.000
5	1.000	.000	.000	.600	.200	.200	.750	.125	.125
6	1.000	.000	.000	1.000	.000	.000	1.000	.000	.000
7	.000	.000	.000	.000	1.000	.000	.000	1.000	.000
8	1.000	.000	.000	1.000	.000	.000	1.000	.000	.000
9	1.000	.000	.000	1.000	.000	.000	1.000	.000	.000
10	1.000	.000	.000	.667	.333	.000	.750	.250	.000
11	.500	.500	.000	.800	.200	.000	.714	.286	.000
12	.833	.167	.000	1.000	.000	.000	.875	.125	.000

Note: Cor = correct decision, Err = erroneous decision, NO = no opinion

A final analysis explored the impact of participant sex on decision accuracy. Table 6 shows total accuracy as a function of participant sex and veracity, using the decisions produced by the blind scorer. There was no evidence for a difference in accuracy by participant sex, *z* = 0.3, *p* > .05. For male participants, there was no evidence for differences as a function of veracity, *z* = 0.6, *p* > .05. However, for female participants, total accuracy was significantly higher for truthful participants than for deceptive participants, *z* = 2.3, *p* < .05. However, this result should be viewed with caution, based on the relatively small sample sizes for female participants in the truthful and deceptive conditions (i.e., n = 18 and n = 11, respectively). This analysis was not conducted for definitive accuracy, because the small incidence of no opinion decisions.

Table 6. Total Accuracy Rates as a Function of Participant Sex and Veracity (Blind Scorer Decisions Only).

	Truthful			Deceptive			Total		
Participant Sex	Cor	Err	NO	Cor	Err	NO	Cor	Err	NO
Frequency									
Female	17	1	0	7	4	0	24	5	0
Male	16	2	0	18	3	1	34	5	1
Proportion									
Female	.944	.056	.000	.636	.364	.061	.828	.172	.000
Male	.889	.111	.000	.818	.136	.046	.850	.125	.025

Note: Cor = correct decision, Err = erroneous decision, NO = no opinion

Discussion

The results of this study showed a high level of reliability (with respect to inter-scorer agreement) and decision accuracy. The accuracy rates produced in this study exceeded the chance levels for truthful, deceptive, and total decisions.

Limitations

As with any research effort, a number of factors must be taken into account when considering the results of this series of studies. First, it should be fully understood that the population from which the samples in this effort were drawn were heterogeneous with respect to sex and race, but were relatively homogenous with respect to age. The average participant in this study was 20. While the race and sex variation exhibited in the sample are of potential value, the restricted age range may limit the generalizeability of the results of this study to other populations.

Second, the present study used mock crime scenario, carefully positioned to capture a potential application of the AFMGQT in the real world. Based on feedback from participants in individual debriefing sessions, the mock crime scenario was engaging and believable. Based on the decision accuracy results, the implications are that the scenario was arousing and engaging. However, concerns over shortcomings of mock crime scenarios in validation studies are well documented, and should be taken into consideration (Iacono, 2000). Though research by Pollina, Dollins, Senter, Krapohl, and Ryan (2004) provides evidence for physiological similarities between laboratory and field-based research, ultimately it is unknown how the AFMGQT would perform with other types of scenarios and operational situations. The degree of generalizeability of the present results to other scenarios and real world contexts is unknown.

Third, the base rate of deception in the present study was approximately 50%. It is likely that in the real world the base rate of deception might be significantly different from this value. Future studies should explore this problem, perhaps using a different base rate of deception, thus capturing a situation that may be more akin to the real world circumstances in which the AFMGQT could be implemented.

Fourth, the present study explored only the use of a two-question AFMGQT, with questions focusing on a specific type of crime. Future studies should also explore the use of this format with three and four relevant questions so that we can expand the knowledge base of the AFMGQT, increasing our understanding of differential effects, if any, as a function of question number and examinee veracity. In addition, other types of simulated scenarios should be explored, using other questions.

Conclusions

This controlled laboratory study provides evidence for the diagnostic value afforded by the AFMGQT format. The AFMGQT produced total and definitive accuracy rates that significantly exceeded chance levels for both truthful and deceptive participants. These results should be taken into consideration in light of the limitations and concerns described above. Additional and continuous research efforts are required to expand the body of knowledge pertaining to the variety of uses afforded by the AFMGQT.

References

Addison, P., Clay, E., & Xie, S. (2003). Worry as a function of public speaking state anxiety type. *Communication Reports*, 16(2), 125-131.

Anderson, P., Rothbaum, B. O., & Hodges, L. F. (2003). Virtual reality exposure in the treatment of social anxiety. *Cognitive & Behavioral Practice*, 10(3), 240-247.

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

Bruning, J. L. & Kintz, B. L. (1997). *Computational handbook of statistics* (4th. ed.). New York, NY: Addison-Wesley (pp. 285-288).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Edition). Hillsdale, NJ: Erlbaum.

Harb, G. C., Eng, W., & Zaider, T. (2003). Behavioral assessment of public-speaking anxiety using a modified version of the Social Performance Rating Scale. *Behaviour Research & Therapy*, 41(11), 1373-1380.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? *Advances in Psychophysiology*, 4, 201-207.

Iacono, W. G. (2000). The detection of deception. In Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (Eds.), *Handbook of Psychophysiology* (2nd Edition). New York, NY: Cambridge University Press, (pp. 772-793).

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

Miller, M., & Ruppert, W. (2004, April). *Replacing Perchlorate in the M115A2 & M116A1 Simulators.* Talk presented at the 30th Environmental and Energy Symposium & Exhibition, San Diego, CA.

National Academy of Sciences (2003). *The Polygraph and Lie Detection.* Washington, DC: National Academies Press.

Pollina, D. A., Dollins, A. B., Senter, S. M., Krapohl, D. J., & Ryan, A. R. (2004). A Comparison of polygraph data obtained from individuals involved in mock crimes and actual criminal investigations. *Journal of Applied Psychology*, 89(6), 1099-1105.

Savitsky, K., & Gilovich, T. (2003). The illusion of transparency and the alleviation of speech anxiety. *Journal of Experimental Social Psychology*, 39(6), 618-625.

UCLA Department of Statistics (2004). *Power Calculator.* http://calculators.stat.ucla.edu/ powercalc/

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine*, 37(5), 360-363.

Zohar, D., Livne, Y., & Fine, J. (2003). The effect of anxiety on linguistic parameters of public speech: A verbal impairment model. *Anxiety, Stress & Coping: An International Journal*, 16(3), 293-306.

Brute-Force Comparison: A Monte Carlo Study of the Objective Scoring System version 3 (OSS-3) and Human Polygraph Scorers

Raymond Nelson[1], Donald J. Krapohl[2], and Mark Handler[3]

Abstract

The authors describe the Objective Scoring System, version 3 (OSS-3) scoring algorithm, and used brute-force statistical methods to compared its accuracy to previously described scoring algorithms and human examiners, using the OSS development sample (N=292) of confirmed single-issue field investigation polygraphs, and a second sample (N=100) of confirmed single-issue investigation cases. OSS-3 demonstrated balanced sensitivity and specificity and provided significant improvements over previous OSS versions in the form of reduced inconclusive results and increased sensitivity to deception. The improvement in specificity to truthfulness was not significant. The Gaussian-Gaussian decision model of OSS-3 was compared to a replication of an empirical Bayesian decision algorithm described by Kircher and Raskin (1988; 2002), and Raskin, Kircher, Honts, and Horowitz, (1988), that was trained on the OSS development sample using discriminate analysis. OSS-3 showed accuracy that met or exceeded that of the empirical Bayesian algorithm. Using Monte Carlo techniques and OSS-3 accuracy exceeded the average decision accuracy of the 10 human scorers, and 9 out of 10 individual scorers, on 6 dimensions of accuracy: overall decision accuracy, inconclusive results, sensitivity to deception, specificity to truthfulness, false negative, and false positive results. Interrater reliability for the 10 human scorers was evaluated using a double bootstrap of Fleiss' kappa, was consistent with previous reported reliability estimates (k = .59, 95% CI = .51 to .66), compared to the expected perfect reliability of the automated algorithm. A cohort of inexperienced polygraph examiner trainees was trained to evaluate the archival sample using a simplified set of scoring rules and optimized decision rules, intended to approximate the function of the new algorithm as reasonable as possible within human scorers. Decision accuracy for the trainees, using the simplified scoring instructions, was not statistically different from that for the experienced examiners. Interrater consistency for the inexperienced scorers was compared to the experienced scorers using a bootstrap resample of the 1000 iterations of the archival sample (N=100). Fleiss' kappa for the student examiner cohort was k = .61, which was not statistically different from the experienced scorers (k = .58). The computer algorithm can be expected to provide perfect reliability. The authors suggest that computer algorithms should be given more weight in quality assurance and field practices, though they caution the need for responsibility surrounding professional opinions and administrative decisions and policies. The authors also encourage further research into the possible development of a simplified rubric for polygraph hand-scoring.

Introduction

Comparison question polygraphy relies on the transformation of physiological reactions to mathematical representations that can be evaluated for empirical classification efficiency or statistical significance. Polygraph scoring research has

[1] Raymond Nelson is the developer of OSS-3 and past contributor to this journal. Requests for reprints should be directed to him at raymond@raymondnelson.us

[2] Donald Krapohl is Past President of the APA (2006-2007) and regular contributor to *Polygraph*.

[3] Mark Handler is a regular contributor to *Polygraph*.

Disclaimer: The OSS-3 algorithm described in this report was developed by the authors to provide an open source, objective and scientifically defensible method for analyzing polygraph data, and was offered openly to the polygraph community. None of the authors has a financial interest in OSS-3.

been almost entirely empirical, with less emphasis on statistical evaluation of the distributions of truthful and deceptive scores. The Objective Scoring system (OSS) (Krapohl & McManus, 1999; Krapohl, 2002) is an exception to this trend, and was designed and normed using the principles of statistical decision making and signal detection theory.

All scoring techniques for comparison question polygraphs involve the transformation of the observation or measurement of differential physiological reactivity to various question types to a numerical index representing the saliency of the target stimulus. In the case of polygraph hand-scoring techniques, data are commonly transformed to ordinal seven-position values, between -3 and +3 (Backster, 1963; 1990; Bell, Raskin, Honts, Kircher, 1999; Handler, 2006; Research Department Staff, 2006; Swinford, 1999) through observation of a test subject's differential reactivity to various test questions within each component sensor. Data are transformed for each presentation of each target stimulus, for each component sensor, and then aggregated to formulate a conclusion.

While the Krapohl and McManus (1999) system is intended to provide a uniform septile distribution of scores, other hand-scoring systems have not completely described the anticipated scoring distributions. There has been some investigation into the frequency of occurrence of numerical values and point assignments (Capps & Ansley, 1992; Krapohl, 1998). Three-position ordinal scales between -1 and +1 have been suggested (Capps & Ansley, 1992b; Van Herk, M., 1990), and investigated (Harwell, 2000; Krapohl, 1998). Blackwell (1998) concluded that seven-position scoring outperformed three-position scoring, but it should be viewed cautiously because there was no adjustment of cutscores for the differences in the distributions of three-position totals compared to seven-position totals.

Polygraph hand-scoring methods vary in their transformation methods. Numerical scoring, as taught at the Defense Academy for Credibility Assessment (Research Department Staff, 2006), requires evaluation of numerical ratios for seven-position score assignments, which imposes requirements for physical measurement of the data.

Some investigators have described rank-order analysis (Gordon, 1999; Gordon & Cochetti, 1987; Honts & Driscoll, 1988; Krapohl, Dutton, & Ryan, 2001; Miritello, 1999). Rank schemes are easily understood nonparametric methods and have shown some promise. However, because ranking replaces the natural variance of the data with a uniform rank variance, rank order schemes may not extend well to scoring systems intended to evaluate multiple simultaneous investigation targets. Efforts to apply rank order schemes to multi-facet or mixed issues examinations can be investigated empirically but will lack both face and construct validity under attempts to reconcile those methods with statistical theory involving the normal variance of differential reactivity to individual investigation targets. Miritello's (1999) description of a rank-order method for mixed question exams is lacking both normative data and a decision model.

Some hand-scoring systems include more features and rules than others. Systems developed by Kircher and Raskin (1988), as described by Bell et al. (1999) and Handler (2006) have systematically reduced interpretable features by excluding those that cannot be reliably and consistently measured, or are not supported by multiple studies. Other differences among the various scoring systems include the interpretation of pneumograph response data before or after the point of answer, interpretation of non-measured criteria such as complexity and changes in respiratory data, and the inclusion of arbitrary numerical data into measurement values when time-domain metrics are described in physical dimension instead of units of time. (see Kircher & Raskin, 2002, and Podlesny & Truslow, 1993, for more discussion regarding this concern.)

Scoring features of the Utah system (Bell, Raskin, Honts, & Kircher, 1999; Handler, 2006), are supported by complete description of their development and validation through discriminate analysis (Kircher & Raskin, 1988; 2002), and are similar to the features described in ASTM standard E-2229-02 (ASTM International 2002) and those currently taught at the Defense Academy for Credibility Assessment (Research Staff, 2006). While CPS (Kircher & Raskin, 1999) and OSS (Krapohl & McManus,

1999) employ features that are familiar to human scorers, PolyScore (Olsen, Harris, & Chiu, 1994; Harris & Olsen, 1994) uses features that were obtained through logistic regression and would be unfamiliar to human examiners. Other available algorithms include Chart Analysis and AXCON (Dollins, Krapohl & Dutton, 2000) in the Axciton computer polygraph (Axciton Systems, Houston TX) and Identifi (Dollins et al., 2000). Those methods employ features, decision models, and normative data that are not completely described in publication.

Despite their differences, polygraph hand-scoring systems are consistent in that greater saliency or differential reactivity to the investigation targets (relevant questions) are correlated with deception. In hand-scoring systems these reactions are assigned negative (-) integer values. Segments of greater differential reactivity to comparison stimuli indicate the investigation targets are less salient or correlated with truthfulness, and are assigned positive (+) integer values.

Hand-scoring values are summed within each test series, for each presentation of each target stimulus, and then summed between test series for each target stimulus. Finally, data for the several target stimuli are summed for a grand total. Field examiners refer to a pair of relevant and comparison stimuli as a *spot*, though the term also applies to the sum of repetitions of each separate investigation target. Polygraph hand-scoring decision policies utilize both total and spot scores, depending on whether the spot scoring rule is used (Capps & Ansley, 1992c). Several studies have investigated the contribution of the spot scoring rule. While previous OSS versions employed decision policies based on total scores, spot scoring has been the predominate method for decision policies pertaining to multiple facets of an alleged incident, or a mixed set of target issues with no known allegations. Although Krapohl and Stern (2003) used the terms *multiple-facet* and *multiple-issue*, the differences between these terms are not universally understood by many field examiners. This can lead to practical and mathematical problems. We will use the expressions *multiple-facet* and *mixed-issue* here to capture the importance of the dependence or independence of the target stimuli. (See Krapohl & Stern, 2003, for a description of multiple-issue testing and the use of combined testing strategies in medical and related testing contexts.)

Various strategies exist to maximize decision accuracy for both multiple-facet and mixed-issue polygraph examinations that are based on a straightforward adjustment of cutscores. Several studies have focused on spot scoring rules, that is, triggering a call of Deception Indicated (DI) based upon strong negative scores to a single relevant question. The use of spot score rules generally improves sensitivity to deception, though at a cost of increased false-positive errors (Senter, 2003; Senter & Dollins, 2002; Senter & Dollins, 2004; Senter, Dollins & Krapohl, 2004). A more careful examination of the underlying statistical distributions would have predicted this effect. It is hardly surprising that we have found no discussion regarding inflated alpha and corrective measures when completing multiple simultaneous significance tests, because polygraph hand-scoring research has generally not investigated the distributions of scores or statistical analysis. Senter's (2003) report that two-stage rules can improve the overall decision accuracy of MGQT exams and provide a more optimal balance of sensitivity and specificity was a procedural solution. It stopped short, however, of statistical procedures such as the Bonferonni correction to alpha, omnibus analysis through the use of ANOVA procedures, or statistical procedures such as the Tukey test, which are designed to manage the complications of multiple comparisons attending to the multiple-facet and mixed-issue examinations. OSS-3 uses Senter's two-stage rules for ZCT and MGQT examinations along with statistical corrections that accommodate the morphology of the underlying distributions.

In our design and laboratory model for the OSS-3 algorithm, we included a Kruskal-Wallis test, as a nonparametric ANOVA, to serve as an omnibus assessment of the significance of differences between the target stimuli of mixed-issues examinations. The present study addresses only the use of OSS-3 with event-specific single-issue polygraphs Zone Comparison Techniques. Additional capabilities of the algorithm will be described in other studies. Design protocols for OSS-3 also include the use of a Test of Proportions, to monitor the distribution of artifacted and uninterpretable values during an examination (Menges, personal communication 3/12/2008). That portion of the OSS-3 algorithm was not evaluated in the present study.

Objective Scoring System, Versions 1 and 2

OSS (Krapohl & McManus, 1999) is based on three simple and mechanically repeatable measurement aspects of polygraph waveforms, called "Kircher features," (Dutton, 2000) which were first described by Kircher & Raskin, (1988). The three Kircher features include: Respiration Line Length (Timm 1982; Krapohl & Dutton, 2001), electrodermal phasic amplitude of increase, and cardiovascular phasic amplitude of increase. Harris, Horner, and McQuarrie (2000) recommended these same physiological indicators as the most robust feature set, and reported these three features as capable of replicating the 7-position numerical scoring system that was in use at the Department of Defense at that time. Kircher, Kristjansson, Gardner, and Webb (2005), provided further argument for this simple set of features as the most robust and reliable feature set for present-day polygraph scoring. These features provide desirable attributes, including that they are easily understood by human examiners or reviewers, are similar to features used in hand-scoring, and they can be mechanically measured with perfect reliability. Both OSS and the Computerized Polygraph System (CPS) (Kircher & Raskin, 1999; 2002) are based on the three Kircher features.

Krapohl (1999) suggested the use of a dimensionless R/C ratio transformation of the Kircher features which became the foundation of earlier OSS versions and was retained in OSS-3. Ratio transformation reduces the mathematical comparison of values to a single ratio value for each measured component sensor, for each presentation of the target stimuli. These ratios are dimensionless in that the physical units of measurement are canceled out algebraically during calculation. R/C ratios are also asymmetrical, in that the distribution of all possible R/C ratios will be a positively skewed distribution of lognormal shape, consisting of all positive real numbers, with a mean of one, an infinite number of possible values between zero and one, and a similarly infinite number of values between one and infinity.

OSS procedures involve the calculation of a physically dimensionless ratio of differential reactivity to various question types, and the transformation of those ratios to a uniform septile distribution of integer values from -3 to +3. OSS total scores are then summed and subject to a Gaussian signal detection model (Wickens, 1991; 2002) that was described by Barland (1985). Krapohl and McManus (1999) provided tables of statistical significance that were constructed using normative data from a large sample of event-specific single-issue investigation polygraphs using Zone Comparison Techniques (ZCT) (Light, 1999) that included three relevant questions concerning a single target allegation, along with three comparison questions and three test series. Dutton (2000) authored a tutorial for the completion of the OSS procedure. Krapohl (2002) provided an update to the OSS normative data, using hand-scoring practices used by examiners trained at the Department of Defense Polygraph Institute; the OSS method remained unchanged at that time.

Krapohl and McManus (1999) reported they satisfied all of their development objectives with the exception of expediency, in that the OSS required some time investment to obtain the physiological measurements. While the earlier OSS required more time to complete compared with traditional pattern recognition approaches to field polygraph hand-scoring, the value of a reliable, well documented, measurement-based, and non-proprietary scoring procedure was not lost, and polygraph instrument developers recognized that deficits in expediency were easily remedied through software automation. The result has been that OSS became a computerized scoring algorithm. Presently three of four manufacturers of computer polygraphs sold in the US have included OSS in their software packages.

Despite its demonstrated efficiency with single issue ZCT polygraph examinations, the practical utility of OSS versions 1 and 2 was limited by the cumulative data structure, and by decision policies that do not attend to the complexities of multi-facet and mixed-issues examinations. The distributions of truthful and deceptive total scores from previous OSS versions are contingent on the number of question presentations regarding a single issue of concern, and on the number of test charts. Total scores are vulnerable to missing or uninterpretable data, as well as to additional data. Therefore, decision norms for earlier OSS versions do not theoretically generalize well to examination techniques involving two or four target stimuli, and are unable to take advantages of the completion of three to five test series as described by Kircher and Raskin (1988), Senter and Dollins (2004), and Senter, Dollins, and Krapohl (2004).

A further limitation of the earlier OSS version is that its data model and decision norms cannot be applied to multi-facet examinations regarding a single known allegation in which the examinee may be truthful to some but not all investigation targets, or mixed-issue screening examinations regarding multiple investigation targets involving unknown incidents. These conditions represent a substantial portion of field polygraph activity, and constitute a need for decision models that can evaluate individual spot scores in addition to total scores.

Krapohl and Norris (2000) evaluated OSS with confirmed criminal investigation exams using the Modified General Question Technique (MGQT, Ansley, 1998; Weaver & Garwood, 1985), and observed that human scorers provided better sensitivity to deception than attempts to apply the total score decision model of the earlier OSS version to spot scoring conditions. Krapohl and Norris also observed that the OSS model outperformed human scorers in terms of specificity to truthfulness. The results of Krapohl and Norris are consistent with mathematical expectations pertaining to the application of a cumulative data model to spot scoring circumstances, in which in the distribution of spot totals, upon which deceptive conclusions are based, can be expected to differ substantially from the distribution of cumulative totals, upon which the OSS-3 method was normed.

Method

Polygraph Component Sensors

Component sensors include upper and lower pneumograph sensors, cardio sensor cuff, and electrodermal sensors. Pulse-oximiter components have been available for some time (Kircher & Raskin, 1988), though they are used less commonly and are not included in presently available computer scoring algorithms. Peripheral activity sensors have become required components in the context of increasingly available strategies intended to defeat the polygraph test. At present, peripheral activity data is not a scored component in hand-scoring or computer algorithms, but is used to confirm the presence or absence of somatic peripheral nervous system activity among the autonomic nervous system data.

Algorithmic Approach

Nelson, Handler, and Krapohl (2007) introduced a major revision to OSS (Krapohl & McManus 1999; Krapohl, 2002), which is now called the Objective Scoring System, version 3 (OSS-3). The data model for OSS-3 is based on the aggregation of data through standardized scores and weighted averaging instead of simple cumulation. The use of standard *z*-scores allows OSS-3 normative data to approximate the distribution of total and spot scores regardless of the number of stimulus targets or test iterations. This important difference makes the OSS-3 method and OSS-3 normative data potentially more widely applicable to a variety of polygraph techniques and polygraph testing circumstances.

The new algorithm uses the mean comparison value as suggested by Elaad (1999), and is similar to previous OSS versions in its use of a two-distribution Gaussian model that was described by Barland (1985). This is in contrast to the single distribution bootstrap algorithm of Honts and Devitt (1992), and the single distribution permutation model of MacLaren and Krapohl (2003). The new algorithm differed substantially from previous versions in its use of standardized values, weighted averaging, and the use of Bootstrap resampling to train normative data for feature standardization, and the two distributions of truthful and deceptive decision norms (see Krapohl, Stern & Bronkema, 2002, for an introduction to probability and distribution models as these concepts apply to polygraph scoring.)

Whereas Krapohl and McManus (1999) managed the asymmetry of R/C ratios through a nonparametric transformation to a uniform distribution of septile bins, we transformed the dimensionless R/C ratios to their equivalent, though symmetrically distributed, natural logarithms. The natural logarithms of the distribution of asymmetrical R/C ratios will become a symmetrical normal distribution with a mean of zero. R/C ratios between zero and one will become an infinite number of logarithmic values between zero and minus-infinity, while values greater than one will become an infinite number of logarithm values between one and infinity. Because lognormal R/C ratios are normally distributed, we are justified in forgoing the granular nonparametric septile transformation of previous OSS versions in favor of parametric statistical procedures that offer greater potential statistical power.

An additional transformation was included at this point. Field polygraph examiners are trained to interpret negative numbers as indicative of greater differential reactivity to target stimuli than to comparison stimuli, and to interpret positive numerical values as indicative of greater differential reactivity to comparison stimuli than investigation targets. Natural logarithms of R/C ratios will be inverse to these expectations. We therefore inverted the sign values of all ratios, so that field examiners who wish to understand the operation of the algorithm can continue to interpret sign values in traditional ways. Sign values for pneumograph data were not inverted, so that human examiners can use a common paradigm for evaluating the data. Data are further transformed by standardizing all values for each component, using normative parameters that were obtained through bootstrap training (Efron, 1982; Mooney 1997; Mooney & Duval, 1993).

Bootstrap training. Bootstrapping is a computer-intensive method of obtaining empirical distribution estimates of parameters such as median and confidence ranges. Under ideal circumstances population parameters such as mean and deviation values would be achieved through testing every member of a population. Because that is often infeasible, test developers depend on samples of data that are intended to be representative of the population on which a test will be used. Variability will always be observed in a sample or population, and it is assumed that some degree of randomness will always be present. As a result, test developers are always concerned about the representativeness of a sampling distribution, and the biasing effect of even small departures from normality. The classical solution to problems of normality and representativeness is to construct numerous sampling distributions from which to calculate the sample parameters, and then use the distribution of sampling distributions as more robust population estimates than could be obtained from a single sample. Modern alternatives to the challenges of constructing numerous sampling distributions involve the use of computer intensive models to gain maximum value from each sampling distribution.

Bootstrapping involves the construction of an empirical bootstrap distribution of resampled sets, with replacement, from the sample data. Resampling is the equivalent of pulling a number at random from a hat after shaking, or randomizing, those numbers and then returning each number to the hat and then re-shaking or re-randomizing the numbers before selecting each subsequent number. This process is repeated continuously to create a resampled distribution of size equal to the sample from which each random selection is drawn. The process of constructing resampled distributions is then repeated numerous times to construct a bootstrap distribution of resampled distributions. With each random case selection, the probability of selecting a case from within the normal range is dictated by the law of large numbers and the central limit theorem, which tell us that if we completed this process a large number of times, our parameter estimates will regress towards the mean of the population represented by the sample.

Bootstrapping can be employed in nonparametric and empirical distribution models and does not depend on normally distributed data. Bootstrapping does assume that sample data are representative of the population, and bootstrapping will not correct for sampling problems. Bootstrap distributions are found to be normally distributed when the underlying sample or population data are normally distributed. Bootstrapping methods can therefore provide robust population estimates for use in parametric statistics, and can also be used to evaluate data for normality.

While it would take a crew of interns several weeks to complete the numerous resampling iterations necessary to achieve bootstrap estimates, modern computers can use brute-force to execute an exhaustive number of iterations with comparative ease. It is not uncommon for bootstrapping experiments to involve 1000 or 10,000, or even more resampled distributions. For each resampled distribution, the statistical parameters of interest are calculated for each resampled set, and a bootstrap distribution of those statistics is constructed by repeating this process many times. It is anticipated that some numbers might be randomly selected more than once within each resampled set, while others may not be selected in all. By using trimmed mean estimates, bootstrap resampling can reduce the influence of outlier or extreme values, against which mean and standard deviation statistics are non-robust or easily influenced.

We created 10,000 resampled sets of size equivalent to the OSS development sample of confirmed ZCT cases (N=292), and calculated population mean and standard deviation estimates for the lognormal R/C ratios for each polygraph component sensor. These values were then used to standardize each of the lognormal R/C ratios in the training sample. Table 1 shows the normative values for the natural logarithms of component ratios, which were derived from the first training bootstrap. Our normative standardization differs from those of Kircher and Raskin (1988; 2002), Raskin, Kircher, Honts, and Horowitz, (1988), who used ipsative standardization of component measurements between all charts to achieve the same goal of algebraically canceling out the physical units of measurement and achieving a consistent metric for evaluating data between the several test charts.

Table 1. Bootstrap mean and standard deviation scores for lognormal R/C ratios by component

	Mean	Standard Deviation
Pneumograph	-0.0385	0.1071
Electrodermal	-0.0179	0.1898
Cardiograph	0.0193	0.4987

Reduction of upper and lower pneumograph data. After standardizing each of the lognormal R/C ratios in the training sample, we then combined the standardized lognormal ratios, for each test stimulus, within each test series. We retained the method of combining upper and lower pneumograph values from previous OSS versions, in which values of opposite numerical sign are set to zero while keeping the signed value of greater magnitude when upper and lower pneumograph values are of similar numerical sign. This method is theoretically capable of retaining more data than the practice of arbitrarily discarding data from one of the pneumograph sensors (Harris & Olsen, 1994), and may be more robust against behaviorally adultered or uninterpretable pneumograph data than averaging the two components.

Trimmed outliers. Before combining the lognormal component ratios within each test chart, we first trimmed all ratios determined to be outliers according to a 3.8906 ipsative standard deviation boundary per each component. This meant that most data values would be considered usable and interpretable, while data values beyond greater than 99.99 percent of other values would be regarded as outliers.

Weighted Averaging. Instead of aggregating the component values through the simple addition methods of previous OSS versions, we combined those values within each test series, for each presentation of each target stimulus through weighted averaging. Several studies have suggested the electrodermal component provides the greatest contribution to diagnostic accuracy (Capps & Ansley, 1992; Harris & Olsen,1994; Kircher & Raskin, 1988; Raskin, Kircher, Honts, & Horowitz, 1988). Kircher, Kristjansson, Gardner, & Webb (2005) showed that cardiograph data is marginally more strongly correlated with the criterion than pneumograph data. Krapohl and McManus (1999) found that weighting the electrodermal component more strongly than cardiograph and pneumograph could reduce inconclusive results without compromising decision accuracy. Earlier OSS versions used integer weighting in which the numerical values assigned to electrodermal data were multiplied by two, meaning that one-half of the total cumulative score from the three component sensors (pneumograph, electrodermal, and cardiovascular) came from the electrodermal channel. We retained the use of integer level weighting, but differed from previous OSS versions in that electrodermal values were multiplied by three, while cardiograph and pneumograph data were multiplied by two and one respectively. The effective result is that component contributions are weighted in the following proportions: electrodermal = .5, cardiograph = .33, and pneumograph = .17. After first aggregating data for each test stimulus within-chart, as just described, we then aggregated the data between the three test series, by averaging the weighted mean scores for each spot. Mean lognormal R/C spot ratios were then further averaged together for a grand mean of standardized lognormal R/C ratios.

Bootstrap decision norms. We completed a second bootstrap of 10,000 resampled sets of the transformed data from the training sample (N = 292), and obtained population mean and standard deviation parameter estimates for separate normative distributions of truthful and deceptive persons for use in a two-distribution Gaussian signal detection model (Wickens, 1991; 2002). Figure 1 displays the quantile-quantile plots which verify that the bootstrap mean and standard deviation estimates for the weighted mean of standardized lognormal ratios of deceptive and truthful subsets are sufficiently normally distributed to justify the use of a parametric z-test in our decision model. Table 2 lists the normative parameters for confirmed truthful and deceptive cases in the training sample.

The ideal way to compare an individual score to those from deceptive or truthful groups would be to have access to the scores of every single deceptive or truthful person. Because that is unrealistic and impractical, test methods are often designed around samples of representative persons from truthful and deceptive groups. If we know how the distribution of scores in those groups (i.e., distribution shape, mean, and variance), then we can use statistical estimates to determine group assignments.

Kircher and Raskin (1988; 2002), and Raskin, Kircher, Honts and Horowitz, (1988) described an empirical Bayesian scoring algorithm that uses maximum likelihood estimates to assign cases to the group which a score most likely belongs to. Another method is the Gaussian-Gaussian signal detection model (Wickens 1991; 2002) described by Barland (1985), in which a score is compared to alternate normative distributions through the use of a simple hypothesis test. OSS-3 is constructed around this method, and assigns a case to the alternate category when the probability is very low, regarding inclusion in one of the normative groups.

The effectiveness of this model depends, in part, on the representativeness of the normative data, an accurate understanding of the distribution shape, and the robustness of our population parameter estimates (i.e., mean, standard deviation). A preferred method of understanding population parameter estimates is to calculate the estimates from a distribution of the mean and variance estimates of numerous sampling distributions, thereby reducing the influence

Figure 1. Quantile-quantile plots for bootstrap mean and standard deviation of weighted mean of standardized lognormal ratios

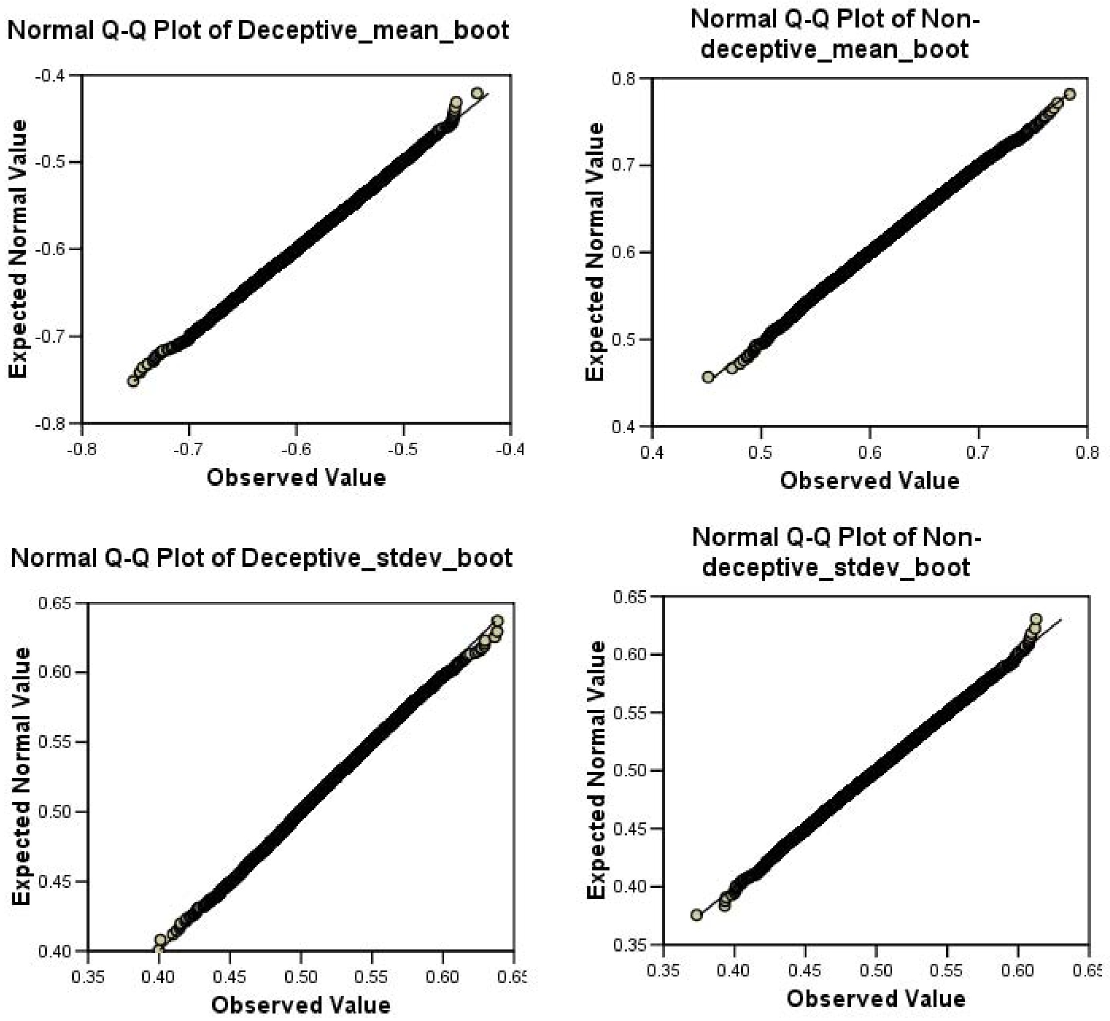


Table 2. Bootstrap mean and standard deviation scores for deceptive and truthful cases in the training sample.

	Mean	Standard Deviation
Deceptive	-0.5863	0.6192
Truthful	0.5188	0.5030

of bias from a single sample. This method depends on access to numerous samples of representative data. A modern alternative is to use brute-force computing and large scale bootstrap resampling methods to reduce the influence of bias in the calculation of population mean and variance estimates.

P-values. Using the bootstrap population norms for the grand mean of standardized lognormal R/C ratios, we evaluated the results of each examination against the distribution of confirmed deceptive cases, using a simple z-test that provides an estimated proportion of confirmed truthful persons that would produce a similar p-value. In field operation, cases are classified as truthful when the resulting p-value is less than the specified decision alpha. Whereas polygraph hand-scoring systems use point totals and cutscores to base decisions, statistical decision models based on signal detection theory make decisions according to alpha thresholds that are analogous to cutscores and represent a predetermined tolerance for risk of error. Very low p-values, when compared to the distribution of confirmed deceptive cases, would alert a field examiner that there is a low probability that the score was produced by a deceptive individual. It is therefore a matter of statistical inference that the individual was most likely truthful regarding the examination target.

Alpha decision cutpoints. Alpha thresholds are matters of administrative policy and tolerance for risk or error, just as much as they are matters of science. Common alpha thresholds are .05, .01, .001, and .1, which represent estimated decision error rates of 1 in 20, 1 in 100, 1 in 1000, and 1 in 10, respectively. Researchers in the social sciences commonly use .05 as a default or arbitrary boundary for statistical significance. Other alpha levels are employed as circumstances warrant. Because OSS-3 uses a two-distribution Gaussian signal detection model (Wickens, 1991; 2002), alpha thresholds for OSS-3 decision must be set for both truthful and deceptive classifications. Because the two alpha boundaries are set independently, they can be set asymmetrically, in order to optimize decision efficiency and balance sensitivity and specificity. Optimal alpha thresholds will served to maximize the correct classification of cases, while constraining inconclusive and erroneous results to acceptable levels. Using data from the training sample (N = 292) we determined that alpha = .1 presented an optimal condition for truthful classifications, including improved specificity to truthfulness and reduced inconclusives, while maintaining a minimal level of false-negative errors.

Two-stage decision policies. By default, OSS-3 uses two-stage decision policies (Senter, 2003) in which truthful classifications are attempted first, followed by attempts to classify cases as deceptive when they cannot be classified as truthful. If a case remains inconclusively resolved after that attempt, a second stage of decision policies is enacted, in which the between-chart mean of standardized lognormal ratios for each spot is assessed. Because the data are combined through averaging and standardization, the distribution and variance of each spot can be approximated by the distributions of grand mean values (see Table 2), which is unaffected by the number of test charts.

When an observed p-value is not less than the specified alpha, compared to the distribution of deceptive individuals, using alpha = .1, the grand mean of standardized lognormal values is then compared to the distribution of confirmed truthful cases, , using alpha = .05 and the same z-test procedure as before. When the resulting p-value is less than the specified decision alpha, it is interpreted as meaning there is a sufficiently low probability the score was produced by a truthful person and a case will be classified as deceptive.

Multiple comparisons and inflated alpha. In the case of single issue ZCT polygraphs, it is inconceivable that a subject could lie to one target stimulus while being truthful to others, or vice versa. Test stimuli are therefore non-independent, or dependent, and the addition rule allows us to calculate inflated alpha levels as $\alpha = \alpha_{\text{per test}}$ x number of tests. This means that while using alpha at .05 for single issue ZCT exams involving three non-independent target stimuli, the inflated alpha level is .05 x 3 = .15. In the case of multiple significance tests that are independent, (i.e., multi-facet or mixed-issues polygraph exams in which it is conceivable that a test subject could lie to one or more target issues while being truthful to one or more other investigation targets), the inflated alpha level can be estimated through the use of the multiplication rule, as $\alpha = 1 - (1\text{-}\alpha_{per\ test})^{number\ of\ tests}$. So, with a multi-facet or mixed-issues polygraph involving three independent targets, the inflated alpha is calculated as $1 - (1 - .05)^3 = .143$. Polygraph exams that use four questions will find the inflated alpha levels even higher. It is important to recognize that polygraph scoring schemes and decision policies based on integer point totals and cutscores are no less immune from multiple comparison and alpha complications. The effects of the addition rule and non-dependency will also play a role in the estimation of the likelihood of inconclusive test results in spot scoring circumstances.

In field polygraph testing, decision policies that neglect to correct for inflated alpha can be expected to contribute to decreased specificity to truthfulness and an increased false-positive error rate. The obvious benefits of completing multiple significance tests and using two-stage rules are the reduction of inconclusive results and improved sensitivity to deception. In the case of multi-facet and mixed issues examinations involving several independent investigation targets, there is also a semantic increase in sensitivity to a broader range of concerns.

A number of statistical and mathematical procedures have been developed to correct for or reduce the impact of inflated alpha levels when completing multiple significance comparisons. The use of a Bonferonni correction to the specified alpha is one of the simplest methods, and applies to both dependent and independent circumstances. Bonferonni correction can be applied to a specified alpha level by multiplying the specified alpha by the number of comparisons. In polygraph spot scoring circumstances involving three stimulus targets, Bonferonni corrected as .05 x 3 = .0167. Senter (2003), and Senter and Dollins (2002) investigated spot scoring and total score decision policies and recommended the adoption of field practices that would serve to manage these known concerns through procedural solutions. It would, however, make equally good sense to begin to describe these concerns using the language of statistical inference that is common to other sciences.

Bonferonni correction. To avoid the increased likelihood of a type-1 error, in the form of false positive results, when completing the second stage of the two-stage scoring rules, we use a Bonferonni corrected alpha during the second stage of the two-stage decision policies. Inflation of the alpha is a known complication in any experimental or testing setting in which multiple simultaneous tests of significance are employed on the same data. With a single test of significance, using alpha at .05, there is a 5% chance (approximately 1 in 20 times) the data will result in a type 1 error and will appear significant due to chance alone. In practice circumstances, type-1 errors are called *false positives*. When conducting multiple simultaneous significance comparisons there is a mathematical inflation of the specified alpha. Calculation of the inflated alpha is typically done by one of two methods, depending on whether the various stimulus targets or investigation issues are independent, or non-independent/dependent.

We found that using two-stage rules (Senter, 2003) improved the sensitivity of the algorithm to deception from .828 to .913, with a corresponding reduction in inconclusive results, from 10.6% to 1.4%. Specificity to truthfulness remained constant at .89. However, the increase in sensitivity was not without cost, as the false positive error rate increased from 1.4% to 10.5%.

The application of a Bonferonni correction to the decision alpha reduced the false-positive rate to 6.5% with a minimal change in sensitivity to .906. Decision accuracy increased with the application of the Bonferonni correction, from 91.3% to 93.9%. To test the significance of these observed differences, we constructed a double-bootstrap Bonferonni t-test. Our double-bootstrap consisted of resampled sets of N = 292 cases from the training sample, from which we calculated mean estimates, before creating an secondary 292 resampled sets for each of the 292 resample sets in the primary bootstrap. The secondary bootstrap was used to calculate variance estimates which we used to complete a series of student's t-tests, using a Bonferonni corrected alpha, due to our use of multiple simultaneous significance tests. Table 3 shows the results of a double-bootstrap Bonferonni t-test. The increase in decision accuracy was significant at *p* < .05, but not significant when compared to the corrected alpha of .008. The reduction in false positive errors was significant, as was the increase in inconclusives. Despite the increase in inconclusive results, the overall inconclusive rate of 4.4% was regarded as tolerable in consideration of the increased decision accuracy and decreased false-positive error rate.

Table 3. OSS-3 (two-stage) results with and without Bonferonni correction.

	Uncorrected alpha	Corrected alpha	sig.
Correct Decisions	.913	.939	.043
Inconclusive	.014	.044	<.001
Sensitivity	.912	.906	.399
Specificity	.889	.889	.483
FN errors	.067	.068	.489
FP errors	.105	.048	.006

Experiment 1

Receiver operating characteristic. Using the receiver operating characteristic (ROC), we calculated the area under the curve (AUC) for OSS-3 and OSS-2. ROC statistics have the advantage of reducing several dimensions of accuracy concern, including sensitivity, specificity, and error rates, to a single numerical value. This makes it possible to easily compare the efficiency of different methods, both numerically and graphically. The advantages of this method become obvious when considering the ease of comparing two numbers compared to that of comparing separate tables of values. Because they evaluate decision accuracy across all possible decision cut-points, ROC statistics can provide analysts and decision makers with estimates of classification efficiency that are more easily integrated into decisions regarding tolerance for risk, compared with the challenges of generalizing accuracy estimates based on tables of values for varying cut-points, or the limitations of a single arbitrarily established cut-point. ROC estimates offer an important advantage over Bayesian estimates in that they are more resistant to base-rate influence. ROC estimates can be thought of as the likelihood that a randomly selected case will be correctly categorized, using a randomly selected decision cut-point. Areas under the curves were AUC = .964 for OSS-3 and AUC = .971 for OSS-2 using the OSS training sample. Data are shown in Figure 2. Table 4 shows that the 95% confidence interval of .945 to .983 for OSS-3 does not differ significantly from the .956 to .987 confidence range observed for OSS-2.

Bonferonni test. We conducted a Bonferonni t-test, using a double-bootstrap distribution of the training sample (N = 292). OSS-3 provided overall performance that equaled or exceed that of OSS-2. The double-bootstrap consisted of 292 samples of the 292 cases in the training sample, for each sample of which we selected an additional 292 samples. Improvements were observed in sensitivity to deception, (*p* < .001), reduced inconclusive results (*p* < .001), and specificity to truthfulness, (*p* = .048) were significant (*p* < .05). Differences in sensitivity to deception (*p* < .001) and reduced inconclusives (*p* < .001) were significant using a Bonferonni corrected

Figure 2. Area under the curve for OSS-3 (two-stage) and OSS-2 (total score).

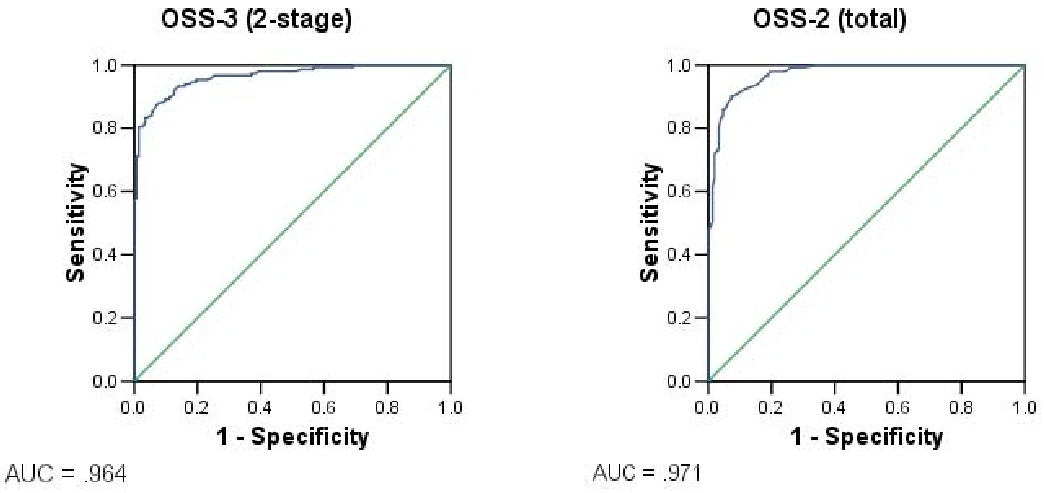


Table 4. AUC for OSS-3 (2-stage rules) and OSS-2 (total score) with training sample (N=292).

	Area	Std. Err.	95% Confidence Interval	
			Lower Bound	Upper Bound
OSS-3 (2-stage rules)	.964	.009	.945	.983
OSS-2 (total)	.971	.008	.956	.987

Table 5. Comparison of performance for OSS-3 and OSS-2 with training sample (N=292).

	OSS-3	OSS-2	sig.
Correct Decisions	93.9%	95.3%	.163
INC	4.4%	12.9%	<.001*
Sensitivity	90.6%	81.9%	<.001*
Specificity	88.8%	84.0%	.048
FN	6.6%	4.7%	.139
FP	4.8%	3.5%	.204

* denotes statistically significant improvement of OSS-3 over OSS-2.

alpha of .008, while the improvement in specificity was not significant at this level. Table 5 shows the results of the bootstrapped Bonferonni test.

Experiment 2

To further evaluate the new algorithm, we replicated the transformations and decision model of the empirical Bayesian algorithm, as described in (Kircher & Raskin, 1988; Kircher & Raskin, 2002; Raskin et al., 1988). That model is based on discriminate analysis and maximum likelihood estimation. Transformations of the empirical Bayesian method algorithm differ slightly from those of OSS-3, in that the empirical Bayesian method uses a *z*-score transformation to achieve a dimensionless measurement of differential reactivity to the test stimulus, whereas the OSS family of algorithms uses an R/C ratio to achieve the same objective (Krapohl, 1999). Transformations differ further in that the empirical Bayesian method uses ipsative standardization for each component, between all charts, and then averages data between charts for each component, before the calculation of maximum likelihood estimates that are then used to make posterior probability adjustments through a Bayesian probability model.

Because it is based on linear discriminate analysis, the empirical Bayesian method combines the between chart means of the component *z*-scores through addition, after weighting those *z*-scores with the unstandardized discriminate coefficients obtained from a discriminate analysis with the training sample (N=292). OSS-3 uses a normative standardization of lognormalized component ratios, and first aggregates data within each chart through weighted averaging of the component scores. Whereas the empirical Bayesian transformations produce a set of mean *z*-scores for all presentations of each test stimulus which are then further averaged for a grand mean *z*-score, OSS-3 transformations produce a weighted mean standardized measurement of differential reactivity for each presentation of each test. OSS-3 transformations then use unweighted averaging to combine the several presentations of each test stimulus for a set mean standard target scores, which are then further averaged for a grand mean score of the standardized lognormal ratios. We used SPSS (version 12.0) to calculate the discriminate function used in our replication of the empirical Bayesian decision algorithm. Table 6 shows the results the unstandardized discriminate coefficients and proportional component weight used in our replication of the empirical Bayesian algorithm.

Table 6. Unstandardized discriminate coefficients and proportional weights.

	Unstandardized discriminate coefficients	Proportional weight
Pneumograph	.629	.192
Electrodermal	1.735	.582
Cardiograph	.920	.280

Table 7 shows that the empirical Bayesian algorithm returned a decision accuracy rate of 94.4% with 7.5% inconclusive results. Sensitivity to deception was .879, while specificity to truthfulness was .865. False negative and false positive error rates were 4.0% and 6.3% respectively.

Table 7. Empirical Bayesian algorithm results with OSS training sample (N=292).

	Probability Analysis
Correct Decisions	94.4%
INC	7.5%
Sensitivity	87.9%
Specificity	86.7%
FN	4.0%
FP	6.3%

In the field of test development, results from a single sample or experiment cannot be regarded as adequately representative of how well a test method will work with the entire population. It is widely understood that accuracy estimates based on development samples are biased or optimistic estimates. Reasons for this include a variety of possibilities which include overfitting of the data model to the sample, reliability constraints with non-automated scoring, and the representativeness of the development sample. In general, simpler data models will not only tend to overfit less often, and will tend to provide greater interrater reliability among human scorers. For these reasons, accuracy estimates with validation samples are regarded as unbiased or less biased estimates.

Data were obtained from an archival sample that was constructed for a replication study conducted by Krapohl and Cushman (2006), which used earlier research (Krapohl, 2005) to develop Evidentiary Decision Rules for manual scoring of examinations conducted using the Zone Comparison Technique (Backster, 1963; Backster, 1990; Department of Defense Research Staff, 2006; Light, 1999).

Evidentiary decision rules (Krapohl, 2005; Krapohl & Cushman, 2006), are useful in field applications such as courtroom and paired-testing, or any testing context in which optimal balance of sensitivity and specificity and minimal inclusive results are among the highest priority. Krapohl and Cushman's sample consisted of N=100 event specific single-issue field polygraph exams, which were selected from an archive of confirmed cases, without regard for the original examiner's opinion. A more complete description of that sample can be found in previous publications. Those examinations were conducted using computerized polygraph systems. After extracting the Kircher features data (Kircher & Raskin, 1988; 2002; Dutton, 2000) using the Extract.exe software program (Harris, in Krapohl & McManus, 1999), we then scored of the replication sample (N=100) using the OSS-3 algorithm and the empirical Bayesian algorithm (Kircher & Raskin, 1988; 2002; Raskin et al., 1988) which were constructed using the open source spreadsheet application OpenOffice.org (available from Sun Microsystems), and a commercial spreadsheet from Microsoft. Artifacted and uninterpretable segments were not included in the computerized scores. Of the 1800 measurements, less than 2% of the data were marked as uninterpretable.

Table 8 shows the results of OSS-3 and the empirical Bayesian algorithm using the replication sample (N=100). Differences in decision accuracy was not significant (*p* = .365), though OSS-3 performed slightly better with 91% correct compared to 90.5% for the empirical Bayesian method. Difference in inconclusive results was significant (*p* = .002) using a Bonferonni corrected alpha of .008, with OSS-3 classifying 6.1% of the archival cases as inconclusive, compared to 15.0% for the empirical Bayesian method. OSS-3 returned fewer false negative errors than the empirical Bayesian method, with 8.1% compared to 12.2%, and more false positive errors, 7.9% compared to 4.0%, though those differences were not significant (*p* = .167) and (*p* = .117) respectively. OSS-3 showed greater sensitivity to deception, .858 compared to .778, which was not significant (*p* = .068). The empirical Bayesian algorithm showed fewer false positive errors, OSS-3 showed better specificity to truthfulness, .861 compared to .761, (*p* = .033) which was significant at .05, but not significant using a Bonferonni corrected alpha of .008. Figure 3 shows the Areas Under the Curve (AUC) for the Receiver Operating Characteristics for OSS-3 and empirical Bayesian algorithm to be .929 and .930 respectively.

To compare the results of the new algorithm to human examiners, we obtained the scored results from 10 human polygraph examiners, who scored the archival sample using evidentiary decision rules described by Krapohl and Cushman (2006). Evidentiary rules use two-stage decision rules (Senter, 2003), and employ cutscores that are empirically shown to reduce inconclusive results and improve specificity to truthfulness compared to traditional decision rules and cutscores.

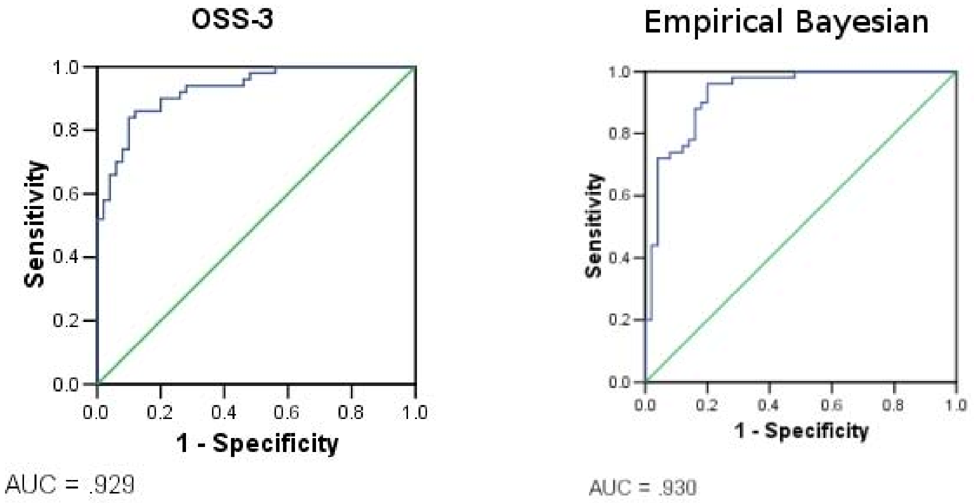
Participants. Human scorers were a self-selected cross-section of field examiners employed in private, law enforcement, and federal polygraph practice. Detailed information was not collected regarding the educational credentials and demographic background of the scorers. Human examiners ranged from 1 to 40 years in experience, with a median of 20 years (mean = 17 years). The examiners volunteered to score polygraph cases to verify their scoring abilities so to qualify for Marin protocol (paired testing) certification (Marin 2000; 2001). Human scorers were permitted to use a variety of existing scoring methods (cf., Backster, 1963; Backster, 1990; Bell, Raskin, Honts, & Kircher, 1999; Department of Defense Research Staff, 2006; Matte, 1996; Matte 1999; Handler, 2006).

Table 8. Comparison of performance for OSS-3 and the empirical Bayesian algorithm with the Marin replication sample (N=100).

	OSS-3	Empirical Bayesian	sig.
Correct Decisions	91.5%	90.5%	.365
INC	6.1%	15.0%	.002*
Sensitivity	85.8%	77.8%	.068
Specificity	86.1%	76.1%	.033
FN	8.1%	12.2%	.167
FP	7.9%	4.0%	.117

* denotes statistically significant difference using Bonferonni corrected alpha = .008.

Figure 3. ROC Area Under the Curve for OSS-3 and the Empirical Bayesian algorithm.



With the Krapohl and Cushman (2006) manual scoring data available we were afforded a benchmark against which to compare the performance of OSS-3. Because the Evidentiary Decision Rules (EDRs) for manual scoring led to the best overall accuracy, all comparisons here used those data with the understanding that the EDRs performance is probably higher than that found in common field practices. Table 9 shows decision accuracy and inconclusive rates for the 10 manual scorers and the OSS-3 algorithm. Decision accuracy rates for the human scorers ranged from 83.3% to 94.6% while inconclusive rates ranged from 4% to 13%. Ranked by decision accuracy, OSS-3 performed as well as or better than 9 of 10 human scorers.

Table 9. Rank order of 10 blind scorers and the OSS-3 algorithm by accuracy, in percent (N=100).

		Correct	Total
Rank	Scorer	Excluding inconclusives	Inconclusive
1	2	94.6	7
2	**OSS-3**	**91.5**	**6**
3	1	89.9	1
4	4	89.7	13
5	9	87.4	5
6	6	86.7	10
7	8	86.7	10
8	5	86.5	11
9	3	83.5	9
10	10	83.5	3
11	7	83.3	4

Experiment 3

We then compared the maximum potential decision accuracy of OSS-3 to the 10 human scorers using ROC analysis. Figure 4 shows that the AUC = .878 for the average of the 10 human scorers in the Krapohl and Cushman (2006) sample. While OSS-3 outperformed the average of human scorers, inspection of the confidence intervals in Table 10 indicate that difference is not statistically significant.

Figure 4. ROC plots for OSS-3 and average of 10 human scorers with replication sample (N=100)

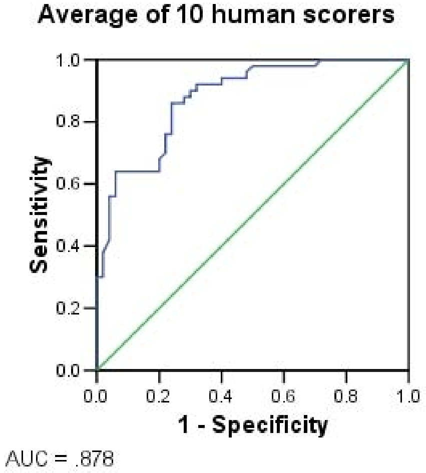


Table 10. AUC for OSS-3 (2-stage rules) and 10 human scores with Krapohl and Cushman (2006) replication sample (N=100).

	Area	Std. Err.	95% Confidence Interval	
			Lower Bound	Upper Bound
OSS-3 (2-stage rules)	.929	.024	.881	.976
10 human scorers	.878	.033	.813	.943

Monte Carlo methods. Next we used Monte Carlo simulation (Mooney 1997; Mooney & Duval, 1993) to compare the accuracy of the OSS-3 to the 10 human scorers. Monte Carlo methods are another class of brute-force computer-intensive methods of simulating the behavior of representative data, using massive sets of random numbers. We began by defining a Monte Carlo population space of N=1000 simulated examinations, for which we used random numbers to assign the confirmation status of each case according to an arbitrary base-rate of 0.5. Next, we use more random numbers to randomly assign the outcome of each deceptive or truthful case according to the proportions specified in Table 11, which are the averaged results of the 10 human scorers in the replication sample. Results for simulated deceptive cases were randomly assigned in the following proportions: correct = 0.792, error = 0.122, and inconclusive = 0.086. Results for simulated truthful cases were assigned according to the following proportions: correct: 0.824, error = 0.116, and inconclusive = 0.060. The use of random numbers for outcome assignment assured that the exact proportion of simulated results would never perfectly conform to the proportions observed in the replication sample, but would vary normally around the specified proportions with each iteration of the Monte Carlo simulation of the population space. Monte Carlo simulation techniques assume that randomness can never be completely eliminated, and uses normal variation in large-scale random simulations to observe how data can be expected to vary in live situations. We used 10,000 iterations of the Monte Carlo space of N=1000 simulated cases to calculate mean estimates and 95% confidence intervals for the 10 human scorers and used those values to calculate the significance of OSS-3 results compared to the human scorers.

OSS-3 produced an overall decision accuracy rate of 91.5% which was significantly better than the 87.2% average decision accuracy of the 10 human scorers (z = -3.91, p <.001). The 6.0% inconclusive results for OSS-3 was not significantly different from the 10 human scorers' average inconclusive rate of 7.2% (z = 1.56, p = .059). Table 12 shows overall decision accuracy and inconclusive rates and 95% confidence intervals from the Monte Carlo simulation, along with the OSS-3 computer algorithm results with the replication sample (N = 100).

Table 11. Averaged results, in percent, for 10 human scorers using Evidentiary Decision Rules.

	Deceptive Cases	Truthful Cases
Correct (with inconclusives)	79.2	82.4
Errors	12.2	11.6
Inconclusive	8.6	6.0
Correct (without inconclusives)	86.2	89.5

Table 12. Overall decision accuracy and (95% confidence interval) for all cases.

	Human Scorers	OSS-3	sig.
Correct (without inconclusives)	87.2 (85.0-89.3)	91.5 ***	<.001
Inconclusive	7.2 (5.7-8.9)	6.0	.059

*** p<.001

OSS-3 outperformed the human scorers with deceptive cases. OSS-3 showed a sensitivity rate of 86.0% which was significantly better than the average sensitivity level of 79.2% for the 10 human scorers (z = -3.75, p <.001), along with fewer false negative errors, 8.0% compared with the 12.2% (z = -2.86, p = .002), and fewer inconclusive results with 6.0%, compared with the 8.6% for the human scorers (z = -2.08, p = .019). Data are shown in Table 13.

Table 13. Decision accuracy and (95% confidence interval) for deceptive cases, in percent.

	Human Scorers	OSS-3	sig.
Sensitivity	79.2 (75.6-82.8)	86.0 ***	<.001
FN Error	12.2 (9.3-15.1)	8.0 **	.002
Inconclusive	8.6 (6.6-11.1)	6.0 *	.019

*** p <.001
** p <.01
* p <.05

The OSS-3 specificity rate of 86.0% was significant compared to the average of 82.4% (79.1% to 85.8%, *z* = 2.1, *p* = .018) for the 10 human scorers in the replication sample. OSS-3 also produced fewer false positive decision errors, with 8.0% compared with 11.6% (8.8% to 14.4%, *z* = -2.49, *p* = .006) for the averaged human scorers. Difference in inconclusive results was not significant for the truthful cases, with OSS-3 returning 6.0% inconclusive results, compared with 6.0% for the averaged human scorers (3.9% to 8.1%, *z* = .006, *p* = .502). Table 14 shows results for truthful cases.

Table 14. Decision accuracy and (95% confidence interval) for truthful cases, in percent.

	Human Scorers	OSS-3	sig.
Specificity	82.4 (79.1-85.8)	86.0 *	.018
FP Error	11.6 (8.8-14.4)	8.0 **	.006
Inconclusive	6.0 (3.9-8.1)	6.0	.502

** p <.01
* p <.05

Experiment 4

To further evaluate the effectiveness of the Kircher features and the possible advantages of a simplified scoring paradigm, we obtained the hand scored results from a cohort of seven inexperienced examiners in their eighth week of training at the Texas Department of Public Safety Polygraph School. These inexperienced examiners, all of whom have previous experience in Law Enforcement, had not yet completed their formal polygraph training and were provided a simplified rubric for polygraph scoring. The simplified hand scoring instructions employed only the three simple Kircher features, and instructions to score the cases in the Krapohl and Cushman (2008) replication sample (N = 100) using three-position scoring.

In an attempt to maximize interrater consistency, those instructions involve one primary scoring rule: the bigger-is-better principle in which any perceptible difference in magnitude between reactions to relevant and comparison stimuli is regarded as a scorable indicator of differential reaction. Two additional scoring guidelines were included in the simplified instructions provided to the inexperienced scorers. First, they were requested to refrain from assigning positive or negative point scores to erratic, artifacted and inconsistent response data, and instead assign a zero value, leave the score as blank or mark the score as artifacted. Second, the inexperienced scorers were instructed to score only those reactions that are timely with the stimulus, avoiding the assignment of negative or positive scores to reactions that occur prior to the stimulus onset or well after the end of the stimulus or point of answer.

Instructions were to score the cases visually, without the aide of measurement devices, using visual discrimination and conservative judgment as the arbiter of ambiguity in the case data. No explicit instructions were provided regarding the length of the scoring window, except a general instruction to score only those reaction segments which they were willing to argue as indicative of a Kircher feature and caused by the stimulus, while not due to artifact feature at the time of the examination. Artifact features include movement distortion, deep breath and other respiratory irregularities, and substantial instability of physiological response data.

Because we were interested in comparing non-mechanical Kircher feature scores with the computer algorithm, special instructions were provided for the interpretation of pneumograph data. Harris, Horner, and McQuarrie (2000) and Kircher, Kristjansson, Gardner, and Webb (2005), reported that Respiration Line Length, (Timm, 1982) provides a reliable approximation to changes in respiratory pattern that were correlated with the criterion of deception or truthfulness. Simplified scoring instructions defined three respiratory patterns as scoreable: 1) increase in respiratory baseline, of three or more respiratory cycles, before return to the pre-stimulus baseline, 2) suppression of respiratory amplitude of three or more respiratory cycles, following the stimulus onset, before return to the pre-stimulus level; and 3) slowing of respiration rate of three or more respiratory cycles from a consistent pre-stimulus level. Instability, movement, deep breaths, holding apnea, and all other features were to be scored zero or marked as an artifacted response segment.

The cohort of inexperienced scorers was instructed to refrain from formulating an opinion or conclusion regarding the truthful or deceptive status of the cases in the replication sample. Instead, we evaluated the mean and variance of the distributions of scores and determined decision cutscores using alpha boundaries common to social science research. Simplified scoring procedures resulted in a mean score of 8.85 for confirmed truthful cases (SD = 7.46), and a mean of -9.63 for confirmed deceptive cases (SD = 8.47). Table 15 depicts those data.

Our earlier experiments informed us that an asymmetrical alpha scheme could reduce the occurrence of inconclusive results among truthful subjects, with little effect on decision errors. We therefore selected scores similar to that used in OSS-3. Deceptive classifications would be made according to an alpha level of α = .05, scored against the distribution of truthful scores, while truthful classifications would be made at α = .1. To avoid an inflation of alpha, and a resulting potential increase in false-positive errors due to multiple statistical comparisons when using spot scores, we used a Bonferonni corrected alpha of α = .0167 for decisions resulting from a single spot. Using data in Table 16, we selected cutscores of +2 (α <= .1) for truthful classifications, and -4 (α <= .5) for deceptive classifications. For deceptive classifications based on spot scores, we used a Bonferonni corrected alpha of α = .05/3 = .017, because the cases in the replication sample included three relevant questions.

Table 15. Mean and standard deviations for truthful and deceptive cases, using the simplified scoring instructions.

	Average	St. Dev.
Confirmed Truthful (N=50)	8.85	7.46
Confirmed Deceptive (N=50)	-9.63	8.47

Table 16. Mean and standard deviations for truthful and deceptive cases, using the simplified scoring instructions.

Distribution of Deceptive Scores		Distribution of Truthful Scores	
NSR Cutscore	Z-value (alpha)	SR Cutscore	Z-value (alpha)
-1	0.154	-8	0.012
0	0.127	**-7**	**0.017**
1	0.104	-6	0.023
2	**0.085**	-5	0.032
3	0.068	**-4**	**0.042**
4	0.053	-3	0.056
5	0.042	-2	0.073
6	0.033	-1	0.093
7	0.025	0	0.118
8	0.019	1	0.146

Table 17 shows the results with the replication sample (N = 100) using the cutscores representing α = .1 for truthful classifications, α = .05 for deceptive classifications, and a Bonferonni corrected alpha of .017 for deceptive classifications based on spot scores obtained using the simplified scoring rubric. These results, obtained from inexperienced scorers, using the Kircher features on which OSS-3 is built, appear to rival those of the experienced scorers reported in Table 11.

We then completed another bootstrap resample of 1000 sets of the replication sample (N = 100), using the data from the 10 experienced using traditional scoring rules (Light, 1999) and seven inexperienced scorers, using the simplified hand-scoring rubric. Table 18 shows there are no significant differences between the results of the experienced scorers, using traditional hand-scoring systems, and inexperienced scorers who used a bare-bones scoring rubric consisting of Kircher features and simple rules.

Table 17. Results obtained with a 3-position hand-scoring rubric (N = 100) using only Kircher features, simplified scoring rules and inexperienced scorers.

	Simplified Hand Scoring
Correct Decisions	87.9%
INC	10.3%
Sensitivity (with inconclusives)	77.4%
Specificity (with inconclusives)	80.3%
Truthful correct (without inconclusives)	85.8%
Deceptive correct (without inconclusives)	90.1%

Table 18. Comparison of experienced scorers (traditional rules) and inexperienced scorers (simplified rules) (N=100).

	Experienced Scorers	Simplified Scoring	sig.
Correct Decisions	86.5%	87.5%	.348
INC	9.6%	10.2%	.416
Sensitivity	80.7%	77.6%	.299
Specificity	75.7%	80.2%	.221
FN	9.5%	12.9%	.225
FP	15.0%	8.9%	.091

To compare differences in inter-scorer consistency between the experienced scorers, and student scorers using a simplified hand-scoring system, we calculated the confidence ranges for Fleiss' kappa statistic for interrater reliability, using a final brute-force computerized statistical analysis, in the form of a two-dimensional double-bootstrap for which both cases and scorers were selected randomly to construct 100 x 100 resampled sets of the replication cases (N = 100). Inter-scorer agreement for the inexperienced scorers using the simplified scoring system (k = .61) had a slight but not significantly better performance advantage (p = .19, ns) over those of the experienced scorers (k = .57) whose reliability coefficient was identical to that reported by Blackwell (1999). Those results are shown in Table 19.

Table 19. Interrater reliability estimates for experienced scorers and inexperienced scorers using a simplified scoring system.

	Fleiss' kappa	95% Confidence Interval
Inexperienced scorers	.61	(.52 - .69)
Experienced scorers	.57	(.50 - .65)

Discussion

These data suggest that OSS-3 is capable of meeting or exceeding the capability of previous OSS versions and many human scorers along several dimensions, including sensitivity to deception, specificity to truthfulness, reduced false-negative and false-positive results, and reduced inconclusive results for deceptive cases. The average of human scorers did not out perform OSS-3 scores on any dimension. Equally important is that the new algorithm is based on mathematical transformations that can be theoretically applied to a much wider variety of examination techniques, including examinations consisting of two to four relevant questions and three to five test charts. The algorithm was designed to accommodate and manage the practical and mathematical complications inherent in multi-facet field investigation polygraphs. Design specifications for OSS-3 include specialized decision policies intended to optimize sensitivity and specificity with mixed-issues screening exams used in law-enforcement pre-employment testing and post-conviction offender testing programs.

We do not recommend increasing the number of investigation targets beyond four relevant questions, though it would be theoretically feasible to do so. Our reasons to advocate constraining the number of acceptable targets are based on the inescapable mathematical compromises necessitated by the effects of common statistical principles for dependent and independent probability events, which advise us to anticipate shifts in error rates that result from the inflation of the specified decision alpha with multiple test questions, and the increase in complex outcomes, including increased inconclusives, resulting from the correction of alpha to levels that would no longer well serve the purposes of field investigation. In short, the addition of more than four independent relevant questions incurs unavoidable errors as well as compromises to the validity of the test results. Constraining the number of investigation targets to four allows a range of flexibility that suits the needs of field polygraph investigators while retaining the ability to manage alpha decision boundaries responsibly.

As always, generalization and external validity of new methods and new knowledge is in part a feature of the representativeness of the normative development and validation sample, and there are known limitations pertaining to the application of polygraph techniques to low-functioning and psychotic persons – both populations which are overrepresented among criminal investigation and forensic subjects. We therefore encourage caution in the use of all polygraph methodologies with all exceptional persons.

The most accurate measure of the effectiveness of any decision model is practical experience in field settings. We conclude that the OSS-3 algorithm is capable of helping to meet the needs of field examiners and researchers, though we caution that the present study was limited to the effectiveness of the algorithm with event-specific/single-issue field investigation cases. Additional research is needed with multi-facet investigative polygraph examinations regarding known allegations, and with mixed-issues screening exams involving multiple investigation targets in the absence of any known allegations.

Although the mathematical transformations and statistical demands of OSS-3 are recognizably more involved than those of previous OSS versions, the procedure can be performed by computers with perfect consistency. Human scorers will never provide reliability that exceeds that of an automated procedure. We have no expectations that field polygraph examiners would attempt to calculate OSS-3 results by hand, but have endeavored to provide a complete description of the OSS-3 method for those who wish to study it. While OSS-3 provides results in the form of recognizable probability values, developers of existing hand-scoring systems in present use have not published or specified any tabular or mathematical methods for the calculation of the level of significance for hand-scored results. Instead existing hand-scoring systems are based on cumulative point totals that pertain to unspecified probability distribution models. Cut-scores for polygraph hand-scoring systems have been investigated for their empirical performance, and may be suboptimal compared with decision thresholds derived through methods based on statistical models. At present, little can be determined regarding how most polygraph cutscores conform to common alpha thresholds in similar signal detection models.

Our ability to study the existing range of computerized scoring algorithms is limited by incomplete documentation for the existing methods, and by proprietary and patent interests that preclude independent investigators, and field examiners from studying and completely understanding those methods. Another difficulty has been the lack of access or difficult access to raw data. We recommend that all manufacturers of polygraph field equipment make data available in a non-binary format that can be easily accessed by researchers equipped with common computer spreadsheets and statistical software. Minimally, all polygraph equipment manufacturers should export the Kircher measurements to a format that is easily machine readable. Presently three companies (Lafayette, Limestone and Stoelting) save these data in an accessible way.

A limitation of all presently available computer based scoring algorithms is that the physiological measurement data cannot be assumed to be robust against artifacts and data of compromised interpretable quality. There is no theoretical rationale suggesting that Kircher features, upon which OSS-3 is built, would be robust against data of marginal or unusual interpretable quality. Similarly, there is no published evidence that any of the features employed by any presently available computer algorithms would robust with uninterpretable data, or can effectively identify data of uninterpretable quality. Present methods of identifying artifacts through extreme values should be regarded as a blunt approach to the problem. Artifacted and uninterpretable data is simply uninterpretable, reminding us of the old adage in computer information processing "garbage in, garbage out." The inclusion of a Test of Proportions in the design specifications for the OSS-3 algorithm does not replace the need for further study in the areas of automated artifact and countermeasure detection. We remind the reader that human examiners should not yet rely on any scoring algorithm without carefully reviewing all test data for interpretable quality.

Research into automated polygraph scoring algorithms began in earnest during the 1980s, and automated algorithms have been available to polygraph examiners since the 1990s. However, there has been a general reluctance among examiners to base polygraph decisions on them. Raskin, Kircher, Honts, and Horowitz (1988) reported that discriminate analysis outperformed blind scorers but did not outperform original examiners. Honts and Amato, (2002) reiterated this conclusion. Honts and Devitt (1992), found no significant differences between the performance of expert human examiners, as original scorers, and the results of two automated algorithms, using discriminate analysis and bootstrapping, and suggested that bootstrapping outperformed the other methods and offered other advantages. Honts and Devitt also noted that their expert examiners were not representative of average field examiners. Research comparing human scorers to other automated algorithms has been mixed. Blackwell (1994), found that early versions of the Polygraph Automated Scoring System (PASS) (Harris &

Olsen, 1994; Olsen, Harris, & Chiu, 1994) did not perform as well in laboratory mock-crime experiments, though accuracy of the computer algorithm appeared to significantly improve with subsequent versions (Blackwell, 1996; 1998). Concerns about the representativeness of the study's human scorers apply to all previous comparisons of computerized and human scorers. The present study is an exception, and includes both experienced and inexperienced human scorers.

Just as the use of brute-force or computer intensive statistical analysis can facilitate our human understanding of the meaning and relevance of obscure physiological signals, the use of automated computer scoring algorithms can foster improvements to human scoring methods and human skills. Presently available computer scoring algorithms may not be capable of considering important nuances in the data as well as human scorers, though Kircher et al. (2005) suggested that original examiners do not seem to benefit from extrapolygraphic information. Nevertheless, there exists a need for further study in the areas of data quality and artifact detection. Standard field practice has been to rely primarily or even exclusively on manual scoring of the polygraph data. Criswell (2007) reported that the American Association of Police Polygraphists has declared it unethical for an examiner to base an opinion solely on the results of a computer scoring algorithm.

Because the mathematical scoring of polygraph test data is concerned only with the identification of statistical significance, we favor the wider use of field practices in which test results are described in terms of *significant reactions* or *no significant reactions* for all types of examinations. Holden (2000) previously discussed the difference between test results and professional opinions. While the presence or absence of statistically significant test results is a matter of objective mathematics, it remains the examiner's responsibility to ensure that nothing other than deception or truthfulness on the part of the examinee would cause those data to appear significant or non-significant. The determination that significant reactions are indicative of deception therefore remains a matter of both professional skill and the accuracy of the psychophysiological constructs that explain why people do or do not respond to polygraph test stimuli (see Handler & Honts, 2008). We do not suggest that a test itself should begin to replace professional responsibility or judgment but rather proffer the concept that algorithmic verification is in fact a useful tool for the field examiner.

Future research should also address the unknown limitations of automated physiological measurement in the presence of artifacted or unusual data quality. Other research should describe the normative distribution of truthful and deceptive scores for various hand-scoring systems, thereby facilitating a more informed statistical comparison of the capabilities of computer-based automated systems against hand-scoring systems. Improved understanding of polygraph hand-scoring norms will assist a variety of scientific investigators to more easily understand and evaluate polygraph decision models. For the present, we recommend further consideration and investigation of the OSS-3 algorithm as a viable scoring system for field use and quality assurance.

In consideration of evidence that OSS-3 and other computer scoring algorithms are capable of outperforming blind human scorers, the results of computer scoring algorithm should be considered carefully in quality assurance activities, though it will remain important for human examiners to review the data for adequacy for automated scoring until algorithms become available to automate those tasks. The use of automated algorithms for quality assurance purposes is less tenable with algorithms or hand-scoring measurements that employ proprietary or idiosyncratic physiological features, as differences between algorithm and human results are far more difficult to understand and resolve. We further recommend further investigation into the merits and possibilities of a simplified hand-scoring system based on Kircher features, simplified scoring guidelines and an empirically justified and statistically based understanding of decision rules and decision cutscores. In consideration of the effectiveness of the three Kircher measurements in both computerized and automated polygraph scoring systems, the use of idiosyncratic features and measurements, for which humans cannot easily understand or for which evidence of feature effectiveness is not available, is not justified.

We do not advocate the surrender of professional judgment to a computer algorithm, or the surrender of professional authority to any test method. Instead we recommend that field examiners, program administrators, and policy makers remain aware that professional judgment and professional ethics are domains of human concern for which there are formidable ethical complications when considering the implications of assigning responsibility for judgment to an automated process. Just as polygraph testing cannot completely substitute for an adequate field investigation, computer algorithms cannot substitute for inadequately administered examinations that suffer from poorly selected examination targets, ineffective linguistic construction, or test data of inadequate interpretable quality. Human judgments and policy decisions may be informed and improved by the results of testing and automated procedures, but the accuracy and effectiveness of those policies and judgment will depend in part on the abilities of those professionals to access complete documentation and data from research. We cannot justify the use of any algorithm with inscrutable features, transformation, decision policies and decision models. As with any evaluation measure, ethical use of a test or automated process requires a reasonable understanding of its design, development goals, and operations, including its strengths and limitations.

References

Ansley, N. (1998). The validity of the modified general question test (MGQT). *Polygraph*, 27, 35-44.

ASTM International (2002). *Standard Practices for Interpretation of Psychophysiological Detection of Deception (Polygraph) Data* (E 2229-02). West Conshohocken, PA: ASTM International.

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

Backster, C. (1990). *Backster zone comparison technique: Chart analysis rules*, Paper presented at the 25th annual seminar of the American Polygraph Association, Louisville, KY.

Barland, G. H. (1985). A method of estimating the accuracy of individual control question polygraph tests. In *Anti-terrorism; forensic science; psychology in police investigations: Proceedings of IDENTA-'85* (142-147). Jerusalem, Israel: The International Congress on Techniques for Criminal Identification.

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

Blackwell, N.J. (1994). *An evaluation of the effectiveness of the Polygraph Automated Scoring System (PASS) in detecting deception in a mock crime analog study*. Department of Defense Polygraph Institute Report DoDPI94-R-0003. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A305755.

Blackwell, N.J. (1996). *PolyScore: A comparison of accuracy*. Department of Defense Polygraph Institute Report DoDPI95-R-0001. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A313520.

Blackwell, N.J. (1998). *PolyScore 3.3 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations*. Department of Defense Polygraph Institute Report DoDPI97-R-006. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A355504/PAA. Reprinted in *Polygraph*, 28, (2) 149-175.

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

Capps, M. H. & Ansley, N. (1992b). Comparison of two scoring scales. *Polygraph*, 21, 39-43.

Capps, M. H. & Ansley, N. (1992c). Analysis of federal polygraph charts by spot and chart total. *Polygraph*, 21, 110-131.

Criswell, E. (2007). Ethics: who really needs 'em anyway. *Police Polygraph Digest*, 1, 6-8.

Dollins, A. B., Krapohl, D. J. & Dutton, D.W. (2000). Computer algorithm comparison. *Polygraph*, 29, 237-247.

Dutton, D. (2000). Guide for performing the objective scoring system. *Polygraph*, 29(2), 177-184.

Department of Defense Research Staff (2006). *Federal Psychophysiological Detection of Deception Examiner Handbook*. Defense Academy for Credibility Assessment (formerly the Department of Defense Polygraph Institute). Ft Jackson, SC. Retrieved 1-10-2007 from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Capital City Press: Montpelier, Vermont.

Elaad, E. (1999). The control question technique: A search for improved decision rules. *Polygraph*, 28, 65-73.

Gordon, N. J. (1999). The academy for scientific investigative training's horizontal scoring system and examiner's algorithm system for chart interpretation. *Polygraph*, 28, 56-64.

Gordon, N. J. & Cochetti, P.M. (1987). The horizontal scoring system. *Polygraph*, 16, 116-125.

Handler, M.D. (2006) The Utah PLC. *Polygraph*, 35(3), 139-148.

Handler, M.D. and Honts C.R. (2008) Psychophysiological mechanisms in deception detection: a theoretical overview. *Polygraph*, 36(4) 221-236.

Harris, J.C., Horner, A., & McQuarrie, D.R. (2000). *An Evaluation of the Criteria Taught by the Department of Defense Polygraph Institute for Interpreting Polygraph Examinations*. Johns Hopkins University, Applied Physics Laboratory. SSD-POR-POR-00-7272.

Harris, J. C., Olsen, Dale, E. (1994). *Polygraph Automated Scoring System*. U.S. Patent Document. Patent Number: 5,327,899.

Harwell, E. M. (2000). A comparison of 3- and 7-position scoring scales with field examinations. *Polygraph*, 29, 195-197.

Holden, E. J. (2000). Pre- and post-conviction polygraph: Building blocks for the future procedures, principles and practices. *Polygraph* , 29, 69-116.

Honts, C. R. & Amato, S.L. (2002). Countermeasures. In Murray Kleiner (Ed.) *Handbook of Polygraph Testing*. (251-264). Academic Press. San Diego.

Honts, C. R. & Devitt, M.K. (1992, August 24). *Bootstrap decision making for polygraph examinations*. Department of Defense Polygraph Institute Report DoDPI92-R-0002. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A304662.

Honts, C. R. & Driscoll, L.N. (1988). A field validity study of rank order scoring system (ROSS) in multiple issue control question tests. *Polygraph*, 17(1), 1-16.

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

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

Kircher, J.C., & Raskin, D.C. (1999). *The Computerized Polygraph System* (Version 3.0) Manual. Salt Lake City, UT: Scientific Assessment Technologies.

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

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

Krapohl, D. J. (1999). Proposed method for scoring electrodermal responses. *Polygraph*, 28, 82-84.

Krapohl, D. (2002). Short Report: An update for the Objective Scoring System. *Polygraph*, 31, 298-302.

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

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

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

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

Krapohl, D. J. & Dutton, D.W. (2001). Respiration line length. *Polygraph*, 30, 56-59.

Krapohl, D. J., Dutton, D. W. & Ryan, A.H. (2001). The rank order scoring system: Replication and extension with field data. *Polygraph*, 30, 172-181.

Krapohl, D. J. & Stern, Brett, A. (2003). Principles of multiple-issue polygraph screening: A model for applicant, post-conviction offender, and counterintelligence testing. *Polygraph*, 32, 201-210.

Krapohl, D. J., Stern, B. A. & Bronkema, Y. (2002). Numerical analysis and wise decisions. *Polygraph*, 32(1), 2-14.

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

MacLaren, V. & Krapohl, D. (2003). Objective assessment of comparison question polygraphy. *Polygraph*, 32, 107-126.

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

Marin, J. (2001). The ASTM exclusionary standard and the APA 'litigation certificate' program. *Polygraph*, 30, 288-293.

Matte, J. A. (1996). *Forensic psychophysiology using the polygraph.* J.A.M. Publications: Williamsville, NY.

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

Miritello, K. (1999). Rank order analysis. *Polygraph*, 28, 74-76.

Mooney, C. Z. (1997). *Monte Carlo Simulation.* Sage Publications: Newbury Park, CA.

Mooney, C. Z. & Duval, R.D. (1993). *Bootstrapping. A nonparametric approach to statistical inference.* Sage Publications: Newbury Park, CA.

Nelson, R., Handler, M. & Krapohl, D. (2007, August). *Development and validation of the Objective Scoring System, version 3.* Poster presentation at the annual meeting of the American Polygraph Association, New Orleans, LA.

Olsen, D. E., Harris, J. C. & Chiu, W.W. (1994). The development of a physiological detection of deception scoring algorithm. *Psychophysiology*, 31, S11.

Podlesny, J. A. & Truslow, C.M. (1993). Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology*, 78, 788-797.

Raskin, D. C., Kircher, J. C., Honts, C. R. & Horowitz, S.W. (1988, May). *A study of the validity of polygraph examinations in criminal investigations*. Final Report, National Institute of Justice, Grant No. 85-IJ-CX-0040.

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

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

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

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

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph*, 28(1), 10-27.

Timm, H. W. (1982). Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. *Journal of Applied Psychology*, 67, 391-400.

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

Weaver, R. S. & Garwood, M. (1985). Comparison of relevant/irrelevant and modified general question technique structures in a split counterintelligence-suitability phase polygraph examination. *Polygraph*, 14, 97-107.

Wickens, T. D. (1991). Maximum-likelihood estimation of a multivariate Gaussian rating model with excluded data. *Journal of Mathematical Psychology*, 36, 213-234.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York: Oxford.

Forensic Interviewing, Polygraph and Child Deception

Stanley M. Slowik[1]

Abstract

This paper discusses some of the problems with the validity and reliability of forensic interviews of young children commonly conducted at Child Advocacy Centers upon which polygraph examinations of accused individuals rely. Specific problems with suggestibility and repetitive interviewing techniques are identified so examiners can recognize inadequate or poorly conducted forensic interviews in cases of child sexual abuse.

Key words: Child Advocacy Center, Forensic Interviewing, Polygraph, Child Deception

Critical to all specific issue polygraph examinations is the investigation that precedes the examination and often both describes the matter being investigated and the person to be examined (Reid & Inbau, 1977). If this investigation is done poorly or terminated prematurely in a rush to carry out a polygraph examination, among other possibilities, the examiner's case fact analysis, strength of issue determinations and relevant and comparison question selection can all be significantly impaired. This, in turn, would likely result in a poorly conducted examination with invalid results. In cases involving the alleged sexual abuse of a child, the pre-polygraph investigation often depends entirely upon the forensic interview of the alleged child victim since medical evidence, corroborating witnesses, pictures or other types of proof of abuse are usually non-existent (Slowik, 2007).

Forensic interviews of children in sexual abuse cases, sadly, have a long history of ineffectiveness resulting in irreparable harm to both the alleged victim and suspect. Traditionally, child interviews were conducted by Child Protection Services, police or Social Services and were done so poorly that both the resulting polygraph examinations of the accused and governmental determinations of guilt were sometimes later proven to be in error (Ceci & Bruck, 1995). Perhaps the most notorious of all of these cases was the McMartin Pre-School case in which Child Protection Services and the police engaged a group of pre-schoolers in intensive and repetitive interviews resulting in 321 criminal charges being filed, a seven year criminal investigation and a trial costing the state of California 16 million dollars (Walker, 2002). Eventually, all charges were dismissed but the lives of the accused – as well as the pre-schoolers – suffered substantial and possibly irreparable harm.

As a direct result of the McMartin and other poorly handled cases, the Child Advocacy Center (CAC) emerged. In theory, professionally trained interviewers at the CAC's were supposed to balance the investigative needs of the police and prosecution while providing the psychological protections of the child therapist and make a determination that there was at least sufficient credibility to the suspected abuse to warrant further investigation if not an arrest. Unfortunately, there appears to be an emerging body of evidence indicating that many current forensic interviews of children

[1] Stanley M. Slowik is a frequent presenter for the APA, the past CEO of John E. Reid & Associates and President of Stanley M. Slowik, Inc., an organization that conducts, creates and offers instruction in various interviewing techniques.

Special appreciation is given to Dr. Frank Horvath who assisted and enhanced the production of this article.

are neither conducted in an objective manner nor possess the diagnostic validity for the determinations they make (Drach, Wientzen & Ricci, 2001; Hagen, 2003). When this is the case, any subsequent polygraph examination of the accused will be directly and negatively affected.

It is therefore the intent of this article to identify some of the more common sources of error associated with the forensic interviews of children, particularly allegations of sexual abuse so that polygraph examiners and others involved in these cases can make assumptions and decisions that are reasonably sound and defensible.

For the purposes of this article, children will refer primarily to pre-schoolers, approximately ages 3 through 6, since the majority of the published research regarding interviews of children allegedly abused uses this specific subject population.

Repetitive Interviewing

There are two basic types of repetitive interviewing situations, both of which have harmful effects upon the forensic interview that polygraph examiners rely upon to construct the polygraph examination of the accused. First there are cases where the child has been interviewed by a number of different people in different settings usually working independently of each other. Second, there are cases where the child is questioned repeatedly by the same agency, sometimes within the same interview session, by the same or different interviewers working in concert with each other.

In the first situation, many adults who talk to children about suspected sexual abuse prior to the forensic interview at the CAC are completely untrained in both investigative and interviewing techniques (Warren & Marsil, 2002). No matter how well intentioned, they say and do things with the child that profoundly and permanently affect subsequent interviews conducted by investigators who have at least received some type of formal training in validated, diagnostic interviewing techniques. These unschooled interviewers typically include parents, relatives such as older siblings, aunts, uncles, grandparents, guardians and neighbors. They also include teachers, medical personnel, coaches, Scout leaders, clergy, nannies and numerous other adults in supervisor positions. Interviews conducted by non-professional interviewers often contain egregious procedural errors such as asking leading questions or instilling bias for or against the accused based on their relationship to the accused (Pool & Lindsay, 1998). This tends to occur more frequently when the outcry takes place during an acrimonious child custody battle (Pool & Lindsey, 1998) and different parties with vested interests all talk to the child before the formal interview at the CAC. Among other serious problems resulting from these often unavoidable early interviews is the introduction of information such as descriptors of sexual anatomy or acts previously unknown to the child that could be misinterpreted as knowledge inappropriate for the child's age and possibly circumstantial confirmation of sexual abuse. It should be noted, however, that research consistently indicates that age inappropriate sexual knowledge and many inappropriate sexual behaviors such as excessive masturbation, public sexual displays, etc. are just as prevalent among children thought never to have been abused as those known to have been abused. Whether introduced by coaching, poor interviewing technique or reasons completely unrelated to the allegation, e.g. seeing pornography on a computer, these symptoms are simply not valid indicators of child sexual abuse (Lanning, 2002). Examiners and others should look for evidence that the forensic interviewer not only actively tried to determine who talked to the child prior to the forensic interview but what specifically was discussed prior to forensic interviewing. Most importantly, the forensic interviewer needs to explore (and the examiner needs to confirm that) certain specific causes of false reports and false memory were specifically considered. Child answers resulting from directed responses, acquired language, threats, promises, bribes, etc. are usually indistinguishable from unmanipulated answers when children are asked open questions such as "What happened?" during the forensic interview. What is often assumed to be the original answer to a question appearing in the interview conducted at the CAC may, in fact, be a changed answer from similar questions that were asked in prior interviews.

In the case of repetitive questions and/or interviews by the same interviewer or agency, many of the same errors cited above can occur, but, as is sometimes the case with interviewers operating independently of each other, two more insidious problems can also occur. Children, and many adults for that matter, often interpret repetitive questioning as an indication that their previous answers have been incorrect (Memon & Vartoukian, 1996). As a result, they often change their answers. Inconsistent and contradictory responses can be a classic symptom of deception since liars often get their stories confused and the stress generated by repeated questioning ("they don't believe me") can cause deceptive people to change their answers to escape this lie-generated stress. One might be able to justify this technique if the investigator could be certain that the initial answers were always deceptive and the "new" answers were now the truth. The problem arises when you consider the non-deceptive motives for changing answers, e.g., a child trying to "please" the interviewer. Unfortunately, since the primary purpose of the initial forensic interview is to determine if there is an adequate basis for the allegation, there is always a substantial possibly that the initial answers were in fact the truth, in which case stress generated by repetitive questioning is actually the cause for either deceptive answers or false memories (Doepke, Hendersen & Critchfield, 2003). It should be noted that in cases where the interviewer has manipulated the child's trust using various rapport building techniques, the child may actually change truthful answers to false responses in an attempt to please the interviewer, maintain the positive emotional attachment with an adult the child admires or to avoid rejection by the interviewer (Russell, 2006). While no ethical investigator would condone badgering or threatening a child as appropriate tactics in child interviews, other actions such as promises, rewards, bribes and overt manipulation of trust ("I'm here to help/protect you, etc.") can be equally effective in creating false responses.

The second insidious effect of repetitive interviewing techniques on both children and adults is habituation to the interviewing process where the same stimuli (questions) do not provoke the same response (answers) when repeated in close temporal proximity (Slowik, 2002). Children and others that have been repeatedly interviewed about the same issue not only become more adept at discovering answers they believe the interviewer wants to hear but tend to display the relaxed, comfortable demeanor more typical of a truthful person – not necessarily because they are being truthful but merely because they've adjusted to the interview process again and again.

For all of these reasons, examiners who must rely upon the forensic interview conducted at the CAC should make sure the interviewer's report indicates a formal attempt to discover any and all interviews regarding the issue under investigation conducted prior to the forensic interview. At the very least, the report should state that the child was specifically asked if they had been questioned by their friends, parents, teachers, relatives, medical staff and other likely adults with a summary of the child's statements included so that previous omissions, inconsistencies and contradictions among interviews become apparent. When it becomes obvious that the forensic interviewer completely failed to ask about prior interviewers, examiners should be extremely cautious about creating the examination based primarily upon the forensic interview and take the time and effort to identify and evaluate the effects of previous interviews and repetitive questioning on the results of the forensic interview.

Suggestibility

One area that researchers and forensic interviewers agree warrant special concern is the problem of suggestibility and young children (Ceci & Bruck, 1993a). Once a child has incorporated something that didn't actually happen into his or her subsequent narrative as to what happened, it becomes virtually impossible to separate real and false memories. Perhaps the most common procedure known to induce false memories in many child interview subjects are the use of props such as anatomically explicit drawings or dolls (Bruck, Ceci, Francouer & Renick, 1995). Some child interview instructional manuals maintain that the use of explicit prompts is a "best practice" and actually advocate the use of such props (Sorenson, Bottoms & Perona, 1997). Nearly all of the published research indicates that, at best, drawings and dolls do not contribute to the accuracy of the child's statements but rather, can lead to false and inaccurate statements resulting from the inherent suggestibility of these devices (Lindberg, Chapman, Samsack, Thomas & Lindberg, 2003). However, it should be noted that if the child has already been specific in describing sexual anatomy and activities prior to the introduction of drawings and dolls, the use of these prompts doesn't appear to create new inaccuracies. Examiners should therefore key to that portion of the forensic interview report in which the child describes the actual sexual abuse. If this was not discussed or revealed before the introduction of sexually suggestive devices, there is a much greater possibility of affordance errors, i.e., statements or activities that occur simply because the child is curious and the interviewer has afforded them a mechanism to act out such as inserting a finger into a doll's orifice only because the doll has an orifice (Fincham, Beach, Moore & Diener, 1994).

Unrelated to suggestibility but associated with the use of prompts in interviewing are errors of omission on the part of the interviewer. If a prop was introduced after a child made allegations of specific sexual acts in response to neutral, open question and the child proceeds to demonstrate something sexual with a prop but the interviewer chooses to ignore the child's actions, the interviewer – intentionally or unintentionally – may be conducting a selective interview. Selective interviews are those in which the interviewer only seeks answers to questions that confirm a pre-determined outcome and fail to explore other unanticipated possibilities (Ceci & Bruck, 1995).

Most of the research on child suggestibility attributes the children's vulnerability to the lack of cognitive development. Therefore suggestibility is very age sensitive with three year olds more likely to make inaccurate statements as a result of suggestion and six year olds less likely during forensic interviews (Quas, Thompson & Clarke-Stewart, 2005). While it is believed that the majority of inaccurate statements made by pre-schoolers are the result of false memories resulting from suggestibility, it should be noted that young children are capable of fabricating lies about serious matters such as sexual abuse even without instruction or coaching (Quas & Clarke-Stewart, 2005). Reporting errors appear to decrease if the child has knowledge or experience regarding the event he or she is reporting. At the same time, research indicates that it's easier to plant false stories of sexual abuse in children who have actually been abused (Pezdek & Hodge, 1999). In most cases, it is assumed that pre-schoolers normally would have no knowledge of sexual anatomy or activity and so it is also assumed that children who make explicit reports of such activity are more likely to be accurate with regard to allegations of abuse. Unfortunately, in today's world, there are other sources of sexual knowledge available to young children besides actually being the victim of abuse. The Internet, adult TV programming, DVD's and other materials are sometimes readily accessible to children who, after viewing, are quite capable of describing very detailed sexual acts. Compounding this problem is the problem of sourcing errors where very young children are sometimes unable to distinguish between things that actually happened to them, things that were told to them by others or things they saw on tapes or in pictures (Walker, 2002). Examiners should be diligent to look for evidence in the forensic interview that the interviewer has conscientiously investigated all of these possible sources of suggestion.

Overall, pre-schoolers are more likely to have more omission errors (failing to report sexual contact that actually occurred), more exaggeration errors (claiming penetration or insertion when they were only touched) and more fabrication errors (claiming to have been touched when they were never touched) than children over six years old (Warren, 2002). Pre-schoolers also tend to be overly inclusive, equating an adult's use of profanity with sexual touching, and they have a difficult time recognizing when they do not understand questions (Warren, 2002). As a result, pre-schoolers tend to answer questions they don't understand and appear to be quite confident in their answers even when they are in error. Ironically, this same phenomenon surfaces with child psychologists, child therapists and other experts with regard to their ability to accurately diagnose children who have actually been sexually abused from children who say they were when they were not. In short, decision confidence does not equate to diagnostic accuracy when it comes to statement accuracy of either purported child victims or child experts (Ceci & Bruck, 1994). Finally, most pre-schoolers simply lack sufficient language and vocabulary skills to articulate certain concepts and events leaving investigators with a paradox; does the child not understand the question or does the child understand but lack the communication skills necessary to respond more adequately (Talwar, Lee, Bala & Lindsay, 2004)? Pre-schoolers tend to be very egocentric and fail to consider the perspective of others. This includes such critical issues as intent (Understanding Child Development, 2001). Thus, they sometimes fail to distinguish between innocent parental touching of genitilia during bathing or medicating and inappropriate sexual touching. Pre-schoolers make more encoding errors, e.g., equating private sexual anatomy with any part of the body covered by clothing. Children tend to have more retention errors where their memories of events are distorted by subsequent events and conversations. Finally, young children tend to have more retrieval errors, again, may actually be due to their inability to articulate their thoughts and recollections rather than not recall an event (Wakefield & Underwager, 1992). Pre-schoolers typically communicate using a vocabulary of only 1,200 to 6,000 words yet as any parent can attest, even infants are very much aware of what is being said and what's happening well before they possess the ability to verbally communicate (Understanding Child Development, APA, 2001).

One of the most critical elements of cognitive development and child deception often inadequately evaluated or completely overlooked by forensic interviewers is the reality that most young children conclude that lying is morally bad because they associate lying (or being caught lying) with punishment (Bussey, 1992). At the same time, they haven't developed the mechanism to view truthfulness as an ethically rewarding experience. In other words, while young children fear being caught lying, they don't see a lot of benefit in being truthful particularly when being truthful results in punishment. Polygraph examiners and interviewers using behavioral techniques that have been validated in terms of diagnostic accuracy have long been aware that even in adults the perception of negative consequences is a critical element in triggering both the verbal and non-verbal responses (Edelstein, Luten, Ekman & Goodman, 2006). Sadly, many forensic interviewing techniques fail to recognize this basic component and actually engage in one-sided procedures that subvert the psychophysiological mechanisms. Rather than manipulate the child's emotional state with stress reducing tactics justified as "rapport building", forensic interviewers should approach the child interview in a neutral, objective fashion and neither attack to increase the child's stress nor pacify to reduce stress. In this fashion observed changes in stress related behavior during the course of the interview are more likely to be the result of the questions and the child's truthful or untruthful response. In general, subjects who appear to exhibit significantly less stress as the interview progresses, particularly if the questions become more sensitive, are more likely to be truthful to the issue under investigation. Conversely, subject's who exhibit a marked increase in behavior associated with stress are more likely to be lying about the event being discussed. Techniques that suggest to the child that "They haven't done anything wrong" or "Nothing will happen to you" artificially bias the behavioral responses. Not surprisingly, if there is little or no fear of detection, both children and adults tend to exhibit more of the behaviors associated with truth tellers, even when lying (Slowik, 2007).

The problems of suggestibility in forensic interviews are most acute when the report of the allegation surfaces while the child is in therapy for some unrelated matter or the child is in therapy at the time of the forensic interview because of the initial report and new allegations and details of abuse surface (Ceci & Bruck, 1993). Some therapeutic methods used with children are particularly susceptible to creating false memories. Hypnosis and inadvertent hypnosis resulting from repetitive questioning has been shown to be likely to create false memories that subjects recall with great confidence as being true (Ofshe & Watters, 1998). Sadly, research indicates that 18% of psychotherapists wrongly believe that people can't lie under hypnosis; some wrongly believe that memories recovered through the use of hypnosis are actually more accurate than other memories and 28% believe that memories of past lives can be recovered (Yapko, 1993). As is the case for repressed memories, there is no valid evidence that past lives, repressed memories or alien abductions actually exist but there is ample evidence that some memory recovery techniques commonly used in therapy can create false memories. Examiners therefore must be especially vigilant when conducting examinations based on reports or forensic interviews generated while the child was participating in therapy. It is important to note that techniques such as hypnosis or the use of anatomically explicit prompts that increase the risk of false memories through suggestibility may in fact be appropriate for therapeutic interviews. Examiners need only be concerned when the initial report or case facts upon which the examination is constructed depend upon information generated by such methods and the accuracy of the child's statements is critical.

Interviewing Technique Errors

Many of the procedures and practices that appear to cause errors in the forensic interviews of children apply equally to interviews of adults where the focus is on the accuracy of the report. When examiners, as part of the normal case fact review and analysis become aware of these problems, the examination should either be postponed until the matter can be successfully resolved or, where possible, procedures adjusted to reduce examination errors. Such would be the case when the examiner discovers that the CAC interview he or she is reviewing is neither the only nor the first interview of the child. In such instances, the examiner would be well served to postpone the interview of the accused until he or she has had a chance to read and evaluate the contents of the child's previous interviews and look for substantive inconsistencies and contradictions. If this is not possible, at the very least, the examiner should carefully question the accused as to any information the accused may have regarding the child's statements made prior to the CAC interview.

One of the most serious problems with many forensic interviews of children involves omissions on the part of the interviewer, specifically, failing to explore alternative possibilities for what the child says happened. Sometimes these omissions are the result of interviewer bias where consciously or unconsciously the forensic interviewer desires a certain outcome. Some forensic interviewing techniques used at CAC's specifically instruct interviewers not to explore alternative possibilities since the child might feel the interviewer is questioning the veracity of the child or the adults who helped bring the complaint forward (Bruck, Ceci & Hembrooke, 1998). While this rationale is questionable even in therapy after the complaint has been investigated and the facts corroborated in so far as these cases can be corroborated, failing to objectively investigate and explore other possibilities is a critical error. Examiners might be able to detect this error by looking for evidence that the forensic interviewer has asked the child who else they've talked to, what was said and if anyone has told the child what to say. As would be the case in any investigation, the child's responses to these questions should be evaluated in terms of significant inconsistencies and contradictions, keeping in mind that poor interviewing techniques alone can cause inconsistencies and contradictions in very young children. Examiners should also confirm that the forensic interviewer has explored motives for deception on the part of the child, either through coaching or on the part of the child acting independently (Ceci, 1993). While the research appears to indicate that the most common cause of inaccurate abuse reports by pre-schoolers is the result of false memories created through suggestibility during poorly conducted interviews, even very young children can fabricate allegations of sexual abuse. Examiners should look for evidence that the more common motives for fabrication (child custody battles, revenge for previous disciplinary actions, need for attention, threats, peer pressure, etc.) were discussed during the formal interview at the CAC when this interview is an important part of the examiner's case facts (Ceci, 1993).

As is the case for all forensic interviews, not just those involving young children, examiners should look for the following errors in technique:

MULTIPLE ISSUE QUESTIONS – "Did he make you touch him while he was touching you?"

COMPOUND QUESTIONS – "Did he touch you on your breasts, butt and between your legs?"

LEADING QUESTIONS – "He didn't make you kiss him, did he?" or "I'll bet he told you not to tell anyone, didn't he?"

DIRECTED RESPONSES – "That's good!", "You don't mean that do you?" or any of a lengthy list of non-verbals such as head nodding/shaking, smiling/frowning, etc. following a child's response.

INTERVIEWER BIAS – Similar to directed responses, but with more emphasis: "Whoever would do things to children is a very bad person and you can help us get him!"

TELEGRAPHING – "People who lie get caught because they can't keep their story together, so you'd better not change anything from what you've said before!" or saying things like "Ya, right!" after a response thought to be deceptive.

MANIPULATION OF THE PSYCHOLOGICAL DYNAMICS – "Now, no matter what, nothing is going to happen to you." Or "I want you to know that you're safe here and you don't have to be afraid of anything" (Lanning, 2002).

As stated previously, most pre-schoolers have not developed the mechanism that creates self-satisfaction in doing the right thing (telling the truth) even when negative consequences result. Pre-schoolers, however, are highly motivated to avoid punishment. While the whole concept of the CAC may have been largely due to overly aggressive interviewing tactics that generated fearful behaviors in truthful subjects, techniques that manipulate the subject's interview using stress reduction techniques wrapped in the guise of rapport building are equally harmful in the opposite direction. Most examiners tend to rely on reports created by the same forensic interviewers at the same CAC's. When this is the case, some effort can be made to periodically review the forensic interviewer's record of outcomes. If the forensic interviewer has enough experience to have created a reasonable body of work where the interview outcomes can be evaluated, examiners should be concerned with excessively lopsided determinations. Depending upon the variety of child sexual abuse cases the forensic interviewer handles and when the forensic interviews are conducted in the investigative process, examiners should look for examples where children's statements were determined to be inaccurate and allegations proven to be false. If, on the other hand, it is always the forensic interviewer's conclusion that there is sufficient evidence to sustain the allegation, with the qualifications stated above, there is a high probability that some of the forensic interviewer's determinations are in error since the probabilities are not historically likely. In fact, the lack of differential probabilities between symptoms that accurately diagnose real from false victims of child sexual abuse is one of the more glaring weaknesses in the interviews commonly promoted by CAC's. Techniques such as the Child Behavior Checklist (CBCL), Symptoms Associated with Sexual Abuse (SASA) and various forms of the Child Sexual Behavior Inventory (CSBI) simply don't have even the minimal accuracy needed to base polygraph examination design decisions upon let alone investigative or court conclusions (Sbraga & O'Donohue, 2003).

Truth Lie Discussions

As the direct result of a number of child interview debacles, the courts, following Federal Rules 601 and 603 have consistently determined that children, with very few exceptions, can no longer be automatically excluded as witnesses simply because they are children and the courtroom experience is presumed to be distressing or emotionally harmful (Lyon & Saywitz, 1999). As reality has proven, testimony by children, including cross-examination, can be conducted in a straightforward manner with no negative effects on the child. In those rare and exceptional cases where there is reason to believe from the pre-trial record that the child fears the accused, courtroom screens and/or video cameras can be utilized. Unfortunately, there are still a number of child developmental experts prepared to argue that all testimony in court by children is innately harmful, that young children are incapable of lying or lying about something as serious as sexual abuse.

In any case, before the child's testimony can be admitted, the child must demonstrate that he/she can identify truth and deception and that they understand the moral implications of lying. Because of this obvious and beneficial legal requirement, forensic interviewers have incorporated the need for this process into the fact-finding interview conducted at the CAC. Some of the more difficult aspects of satisfying these legal requirements are the undeveloped cognitive and language skills present in most pre-schoolers. Thus the child may actually have an acceptable understanding of lying and the consequences of lying but be simply unable to express him or herself. Ideally, it would be desirable for interviewers to just ask open questions such as "What is a lie?" or "What happens to people who lie?" and then evaluate the child's abilities and understandings. In an attempt to prevent the exclusion of children's testimony in cases where a child possesses the requisite cognitive ability but lacks the language skills to explain it, forensic interviewers sometimes ask more specific questions or present scenarios to demonstrate the child's understanding. Unfortunately, this sometimes results in a meaningless discussion of color or pet identity completely subverting the legal requirements (Huffman, Warren & Larson, 1999). For example, the forensic interviewer might hold up a green crayon and ask the child "If I tell you this crayon is red, is it a lie?" Virtually all authorative definitions of deception require consideration of intent (Strichartz & Burton, 1990). Thus, if someone were merely mistaken because they don't know the name of a color or pet species, they would not be considered liars. Compounding the problem, there are interviewing manuals recommending tactics like the "color test" as a best practice with absolutely no research to indicate that children who know their colors understand deception or the morality of deception any better than children who "fail" the color identification test. Polygraph testing itself is not immune from the reality that the common practice is not necessarily the best practice (Krapohl, Stern & Ryan, 2003). Ironically, even when forensic interviewers use techniques that have actually been validated to prove the child can tell the difference between truth and deception and understands the moral consequences of lying, this ability is no guarantee that the child will tell the truth (Gilstrap & McHenry, 2006). As is the case with adults, people who lie about criminal activity do so primarily to avoid negative consequences which can result both from being caught lying and from being truthful about wrongdoing. Finally, there are those who recommend – as a Best Practice, no less - subverting the investigative interview and the FRE 601 and 603 requirements by interjecting procedures that might be appropriate during therapy but either bias the child's responses or contaminate the interview so it cannot be introduced as part of any formal legal proceeding (London & Nunez, 2002). These tactics include deliberately incorporating therapeutic procedures into the fact-finding interview so the session can resist review and discovery by claiming to be privileged therapy and not part of an investigation. Other tactics include failing to tape the interview, tapes of such poor quality that they cannot be transcribed or reviewed, intentionally telling the child that the interviewer may repeat some questions knowing the problems of repetitive questioning and suggestibility or actually stating that the tape will or cannot be used in court.

Again, it should be noted that most of these concerns are only relevant to the basic fact-finding interview upon which the polygraph examination relies. If child therapists can demonstrate that manipulative tactics have a beneficial impact in therapy, they should not be held to the same standards of practice required in validated investigative or forensic interview.

Interview Training

One of the most important findings of the research involving forensic interviewing is the relationship between the length and type of interview training and interviewer's competency with regard to obtaining accurate information from children. Three elements should be present to ensure an interview that can be relied upon; use of an interviewing technique that has been validated for diagnostic accuracy; a lengthy initial course of instruction that includes a captive internship at least some of which involves real-life subjects in actual sexual abuse cases; and, some type of quality control where actual case interviews are periodically reviewed and critiqued.

Only one state today (Illinois) requires that student polygraph examiners undergo a captive internship in which they must conduct real-life cases under direct supervision of a licensed instructor and most states do not even license polygraph examiners. A similar situation exists for students of forensic interviewing (Russell, 2006). As was the case for medicine, law and other professions, students who learn by direct observation and discussion – in addition to formal academic instruction - tend to be more proficient in the real world application of their knowledge. The flaw with this system, of course, is that it entirely depends upon the competency and willingness of the mentor to teach the student the correct procedures, an assumption often proven to be false. Starting with Reid College, polygraph training schools slowly accepted that their basic techniques need to be validated as diagnostically accurate and their teaching methods proven effective (Horvath, 2007). Sadly, there are still teaching institutions within the polygraph profession that have never established the validity or reliability of the techniques they teach.

Forensic interviewing is still in the early stages of the validation process with regard to the accuracy of information obtained. It appears that many forensic interviewers today use techniques, while purported to be accepted or even best practices that have never been scientifically researched with regard to the accuracy of information obtained. Some promote interviewing procedures such as the use of drawings and dolls that do not increase the quality of information but can actually increase inaccurate and false information because of suggestibility. Polygraph examinations are dependent upon the investigations, which precedes them which, in the case of child sexual abuse allegations, are heavily dependent upon CAC interviews. Examiners should make an effort to discover how many interviews of the child of what type were conducted by whom using what techniques and carefully evaluate the accuracy of information. Particular attention should be made to evaluate with the effects of repetitive and suggestive interviewing practices. Inadequate or inaccurate case facts will almost always result in an inaccurate polygraph examination.

References

American Psychological Association (2001). *Understanding Child Development as a Violence Prevention Tool.* American Psychological Association, Washington, DC, 1-20.

Bruck, M., Ceci, S., Francouer, E. and Renick, A. (1995). Anatomically detailed dolls do not facilitate preschools' reports of a pediatric examination involving genital touching. *Journal of Experimental Psychology: Applied*, 1, 95-109.

Bruck, K., Ceci, S. and Hembrooke, S. (1998). Reliability and credibility of young children's reports. *American Psychologist,* 53, 136-151.

Bussey, K. (1992). Lying and truthfulness: Children's definitions, standards and evaluative reactions. *Child Development*, 63, 129-137.

Ceci, S. and Bruck, M. (1993a). Suggestibility of the child witness – A historical review and synthesis. *Psychological Bulletin*, 113, 403-439.

Ceci, S. and Bruck, M. (1993b). Child witnesses: Translating research into policy. *Social Policy Report,* 7, 1-30.

Ceci, S.J., and Bruck, M. (1994). How reliable are children's statements? …It depends. *Family Relations*, 43, 255-57.

Ceci, S.J., and Bruck, M. (1995). Children's allegations of sexual abuse: Forensic and scientific issues: A reply to commentators. *Psychology, Public Policy and Law*, 1, 494-520.

Drach, K., Wientzen, J. and Ricci, L. (2001). The diagnostic utility of sexual behavior problems in diagnosing sexual abuse in a forensic child abuse evaluation clinic. *Child Abuse and Neglect*, 25, 489-503.

Doepke, K., Henderson, A., and Critchfield, T. (2003). Social antecedents of children's eyewitness testimony: A single-subject experimental analysis. *Journal of Applied Behavior Analysis*, 36, 459-463.

Edelstein, R., Luten, T., Ekman, P., and Goodman, G., (2006). Detecting lies in children and adults. *Law and Human Behavior*, 30, 1-10.

Fincham, F.D., Beach, S.R., Moore, T., and Diener, C. (1994). The professional response to child sexual abuse: Whose interests are served? *Family Relations*, 43, 244-54.

Gilstrap, L. and McHenry, M. (2006). Using experts to aid jurors in assessing child witness credibility. *The Colorado Lawyer*, 35, 65-74.

Hagen, M. (2003). Faith in the model and resistance to research. *Clinical Psychology: Science and Practice*, 10, 344-348.

Horvath, F. (2007). A review and critique of Alder's "The Lie Detectors, The History of an American Obsession": What polygraph examiners should know. *Polygraph*, 36, 211-219.

Huffman, M. Warren and Larson, S. (1999). Discussing truth and lies in interviews with children: Whether, why and how. *Applied Developmental Science*, 3, 6-15.

Krapohl, D. Stern, B. and Ryan, A. (2003). Exclusionary or nonexclusionary: A review of the evidence. *Polygraph*, 32, 245-250.

Lanning, K. (2002). Criminal investigation of sexual victimization of children. *The APSAC Handbook on Child Maltreatment, 2nd Ed.*, Thousand Oaks, CA, London, New Delhi, Saga Publications.

Lindberg, M., Chapman, M., Samsack, D. Thomas, S. and Lindberg, A. (2003). Comparisons of three different investigative interview techniques with young children. *Journal of Genetic Psychology*, 164, 5-28.

London, K., and Nunez, N. (2002). Examining the efficacy of truth/lie discussions in prediction and increasing the veracity of children's reports. *Journal of Experimental Child Psychology*, 83, 131-147.

Lyon, T., and Saywitz, K. (1999). Young maltreated children's competence to take the oath. *Applied Developmental Science*, 3, 16-27.

Memon, A., and Vartoukian, R. (1996). The effects of repeated questioning on young children's eyewitness testimony. *British Journal of Psychology*, 87, 403-415.

Ofshe, R., and Watters, E, (1998). *Making Monsters: Child Sexual Abuse and False Memory Syndrome*. Amherst, NY, Prometheus Books, 227-249.

Pezdek, K. and Hodge, D. (1999). Planting false childhood memories in children: The role of event plausibility, *Child Development*, 70, 887-895.

Poole, D. and Lindsay, D. (1998). Assessing the accuracy of young children's reports: Lessons from the investigation of child sexual abuse. *Applied and Preventive Psychology*, 7, 1-26.

Quas, J., Thompson, W., Clarke-Stewart, K., (2005). Do jurors "know" what isn't so about child witnesses? *Law and Human Behavior*, 29, 425-456.

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

Russell, A. (2006) Best practices in child forensic interviews: Interview instructions and truth-lie discussions. *Public Law and Policy*, 28, 99-13.

Sbraga, T. and O'Donohue, W. (2003). Post hoc reasoning in possible cases of child sexual abuse: Symptoms of inconclusive origins. *Clinical Psychology: Science and Practice*, 10, 320-34.

Slowik, S. (2002). *The Event Analysis Interview*. Presented at the American Association of School Personnel Administrators, 64th Annual Conference, Vancouver, BC.

Slowik, S. (2007), *Child Deception and Forensic Interviewing*. Presentation at the American Polygraph Association Annual Seminar, New Orleans, LA, 1-20.

Sorenson, E., Bottoms, B. and Perona, A. (1997). *Handbook on Intake and Forensic Interviewing in the Children's Advocacy Center Setting, Office of Juvenile Justice and Delinquency Prevention and the National Network of Children's Advocacy Centers*, p.1-103.

Strichartz, A., and Burton, R., (1990). Lies and truth: A study of the development of the concept. *Child Development*, 61, 211-220.

Talwar, V., Lee, K., Bala, N., and Lindsay, R. (2004). Children's lie-telling to conceal a parent's transgression: Legal implications. *Law and Human Behavior*, 28, 411-443.

Wakefield, H., Underwager, R., (1992), Recovered memories of alleged sexual abuse: lawsuits against parents. *Behavioral Sciences and the Law*, 10, 483-507.

Walker, N., (2002). Forensic interviews of children: The components of scientific validity and legal admissibility. *Law and Contemporary Problems*, 65, 149-180.

Warren, A., Marsil, D. (2002). Why children's suggestibility remains a serious concern. *Law and Contemporary Problems*, 65, 127-147.

Yapko, M.D. (1993). The seductions of memory. *Family Therapy Networker*, 3, 31-37.

The Concept of Allostasis in Polygraph Testing

Mark Handler, Louis Rovner and Raymond Nelson

Abstract

This paper introduces the polygraph profession to the concept of *allostasis*, as a model of physiologic regulation. We compare here the regulatory models of allostasis and homeostasis as potential causes of differential arousal as measured in field polygraphy. In polygraphy, the term homeostatsis is often incorrectly used to describe a waveform observed when an examinee's physiological condition is stable. Allodynamic regulation (Berntson & Cacioppo, 2007) is a result of integrative processes occurring within the central nervous system and mediated by the neuroendocrine systems (Janig, 2006). This concept of regulation is proposed to describe a portion of the physiological arousal observed during polygraph testing.

Classic Homeostasis

Homeostasis is a term used within the scientific community to describe the maintenance of the internal viability of organisms (Schulkin, 2003). The word homeostasis is derived from the Greek homeo, means "same," while stasis means "stable;" thus, "remaining stable by staying the same." Cannon (1932) coined the term "homeostasis" to refer to the processes by which constancy of the fluid matrix is maintained (Berntson & Cacioppo, 2007). Claude Bernard (1878) declared "All the vital mechanisms have only one object, to preserve constant the condition of the internal environment." Studies in physiology and medicine have interpreted that statement to mean certain aspects of the internal milieu are clamped or fixed at a specific setpoint. The historical concept of homeostasis is the basis of modern concepts of autonomic regulation and control (Berntson & Cacioppo, 2007).

Much like a thermostat in a home, homeostatic reflexes adjust to maintain a constant setpoint or level. Homeostasis involves what is called a negative feedback loop because it waits for something to happen before acting. A feedback loop involves a central control module which receives input regarding a condition, processes it and sends an output signal to maintain a setpoint. The central control center in a negative feedback system sends a correction to reverse the change from a setpoint to maintain a constant or fixed state. Positive control feedback systems enhance a stimulus that is already present. The classic feedback control model of homeostasis in psychophysiology describes compensatory responses to restore detected imbalances rather than enhancing what is already there (Berntson & Cacioppo, 2007) and thus is considered negative. Homeostasis describes the regulation of the body to a balance, by single point tuning such as blood pressure, blood oxygen level, blood glucose or blood pH. Baroreptor reflex in blood pressure is the classic, prototypic homeostatic system whose inputs, outputs and controls are well characterized. But blood pressure setpoints can, and do, change depending on the circumstances. Additionally, blood pressure can be changed through a variety of ways, not necessarily through one simple negative feedback system.

Acknowledgements

The authors are grateful to Dr. Gary Berntson, Dr. Michelle Grenier, Dr. Tim Weber, Dr. Peter Sterling, James R. Wygant, Dr. Charles Honts and Donald Krapohl for their thoughtful reviews and comments to earlier drafts of this paper. The views expressed in this article are solely those of the authors, and do not necessarily represent those of the American Association of Police Polygraphists or the American Polygraph Association. Questions and comments are welcome at polygraphmark@gmail.com.

Allostasis

The use of the term allostasis more accurately describes the physiologic mechanisms at work during polygraph testing. Sterling and Eyer (1988) introduced the term allostasis to describe the complexities of visceral regulation (Berntson & Cacioppo 2007), suggesting the model of homeostasis was insufficient to describe the phenomena of changing physiological parameters to meet challenges. While some physiologic regulators maintain a stable setpoint, this is not the case with all physiologic regulation. Mean values of certain parameters are not necessarily fixed points, but rather a setpoint most frequently demanded (Sterling, 2004).

Allostasis is the process of achieving stability, or homeostasis, through physiologic or behavioral change. This term is derived from the Greek: *allo* meaning change, and *stasis* meaning "stable." That is, some changes are necessary to maintain stability or viability. These changes are presumed to be aimed at ensuring the overall viability of the organism. Allostasis encompasses both behavioral and physiologic processes directed towards maintaining adaptive states of the internal environment. One common example is the ever changing relative blood pressure in a person over the course of the day. Researchers have found mean arterial blood pressure fluctuates to meet demands or in an anticipation of a demand (Bevan, Honour, & Stott, 1969).

Therefore, a person in an allostatic state will seek refuge or recovery from that state once the condition causing that state has passed. It is not proper to assert they are out of homeostasis during periods of reaction to the test stimuli. Certain physiologic system parameters of polygraph test subjects seek homeostasis, despite the impositions of stress or stimuli. Reactions during polygraph examinations are an allodynamic regulatory attempt to maintain homeostasis in response to the application of the stimulus. What is commonly referred by polygraph examiners as "relaxation," "recovery" or "relief" is also an allodynamic adjustment to maintain homeostasis with the passing of the stimulus. Allodynamic regulatory systems are not switched on but are always present and respond to help the organism adapt to changes in environmental stimuli or demands.

The allostatic model acknowledges the organism can use prior information to predict demand and adjust proactively before the demand is needed. Cannon (1928) recognized the body can respond in anticipation of a disturbance or agitation. For example, blood pressure typically rises slightly during the moments just before a person stands after sitting or relaxing. The anticipatory increase in blood pressure is adaptive, and serves to prevent lightheadedness by preventing the gravitational pull of blood to the feet by this positional change. The anticipatory increase in blood pressure is not in response to environmental or physiologic feedback, but can be thought of as a form of adaptive learning from past experiences with the action of standing (Dworkin, 1993). If a subject takes medication which blocks these blood pressure changes, the feed forward action can be blocked and the subject becomes dizzy.

Emotionality and Allodynamic Regulation

Emotionality can be used to describe a response based on the perceived value of a stimulus and may include such things as fear and anxiety. Generally fear is an emotional reaction to a present and threatening stimulus, from which the organism seeks refuge, relief or escape, and anxiety is concern of *what might happen* (Le Doux, 2002). Both are an adaptive response, rooted in our evolutionary past, regulated by neuroendocrine events that control behavioral and autonomic responses (Schulkin, 2003). The emotion of fear is dependent upon the neural activity of the amygdala (a small walnut-shaped part of the brain located in the anterior pole of the temporal lobe and being part of the limbic system). Parts of the amygdala have been called the sensory gateway as they receive information from numerous processes of the brain (Aggleton & Mishkin, 1986). The amygdala has been associated with our ability to predict fear (Sterling 2004), and damage to the amygdala has been linked to a reduction in fear-related responses (Le Doux, 1996). Anxiety has been described to be associated with the septal-hippocampus (Gray & Mc Naughton, 2003) or the bed nucleus of the stria terminalis, a portion of the extended amygdala (Heimer, Van Hoesen, Trimble & Zahm, 2008; Walker, Toufexis & Davis, 2003). Because any number of things may contribute to the underlying cause of reaction during polygraph examinations, we will use the term emotionality to more safely subsume the potential causes.

There is no clear setpoint for any particular emotion and thus it is better described under the concept of allostasis than homeostasis. Allostasis describes the changes that occur behaviorally and physiologically to facilitate survival based on an assessment of the stimulus. Once the dangerous condition has passed and the organism experiences relief, the arousal state should subside, and allodynamic regulation should function to restore setpoints.

Allostatic load is a term used to describe the wear and tear on the body as a result of psychophysiologic change. McEwan and Wingfield (2003) propose two types, Type 1 and Type 2 allostatic loads which result in different responses. For each system of the body, there are both short-term adaptive actions (allodynamic regulations) that are protective, and long-term effects that can be damaging (allostatic load).

Type 1 allostatic loads occur when energy demand exceeds supply. The organism moves into a survival mode in an effort to mitigate allostatic load. Once the emergency has passed, the animal returns to a “normal” level of existence. Type 1 allostatic loads are those we are likely to encounter during polygraph testing.

Type 2 allostatic overload begins when there is sufficient or even excess energy consumption accompanied by social conflict and other types of social dysfunction. Type 2 allostatic overload does not trigger an escape response, and can only be counteracted through learning and changes in the social structure.

Arousal through the Emotional and Motivational Path during Polygraph Testing

The exact nature of emotionality underlying arousal during polygraph testing may not be known, may vary by test subject and may present a lofty challenge to “tease out” in scientific testing. There does, however, seem to be face validity around the idea that some degree of emotionality may be present during field examinations. Davis (1961) provided three possible explanations for reactions during polygraph testing. These include the theories of: conditioned reactions, fear of punishment and conflict. All of these are based on an emotional or motivational component as the underlying cause of arousal which he may have linked to fear. The conditioned reaction theory states that involvement in the issue under investigation has created a learned or conditioned response potential, through the action of classical conditioning. A polygraph question becomes a conditioned stimulus and response magnitude may be commensurate with the amount of salience that stimulus holds for that examinee. When the examiner discusses the crime with a guilty examinee during the pretest interview, the contextual recall will result in autonomic arousal. The fear of punishment (fear of consequences) theory postulates a guilty examinee will experience autonomic arousal as a result of fear of consequences of discovery or false accusation. The conflict theory suggests a “guilty” examinee will experience arousal due to internal conflict arising from the motivational forces that cause him or her to answer the questions falsely.

The Defense Academy for Credibility Assessment (DACA), formerly the Department of Defense Polygraph Institute (DoDPI) Anatomy and Physiology for the Forensic Psychophysiologist chapter (DoDPI, 1994) handout states the reactions we expect (or hope to see) during a polygraph examination result from fight, flight or freeze reactions. These include increases in blood pressure, heart rate increase, increase in the contractile force of the heart, redistribution of blood in the body, increase in skin conductance, decrease in skin resistance, dilation of the bronchi and faster deeper breathing.

Many of the physiologic changes reported to occur during fight, flight or freeze can account for changes we see in polygraph tracings following the presentations of a stimulus (test question). Arousal during polygraph testing may be due to fear, stress, guilt, anger, excitement or an examinee’s orienting response to information (National Research Council, 2003).

Lying is an avoidance reaction that can induce arousal through anxiety, stress or guilt. Motivation can increase arousal and in polygraph testing motivation can be great if the consequences of failing are serious (Gustafson & Orne, 1963). It is clear that measurable physiologic reactions occur in response to polygraph questions, and that a number of psychological processes are related to those physiologic reactions, including conditioned response, anxiety, fear, conflict, complexity and other phenomena. The degree to which each may contribute to the allostatic state remains unknown and hence our use of the term emotionality.

Conclusion

The concept of allostasis and allodynamic regulation is not incongruent with the longstanding model of homeostasis. Allodynamic regulation is a conceptual expansion of a single setpoint model and well describes a multi-systemic response aimed at adjusting internal setpoints to meet the demands of the moment. The concept of homeostasis is grounded in the idea of a single optimal level for any given bodily measure. Optimal levels may change at any given moment, based on the current or anticipated circumstances. The marvel of the concept of allostasis (and allodynamic regulation) is it describes how the body can prepare for an inevitable change in any setpoint without having to wait for it to happen. Allostasis takes into consideration the idea that the central nervous system accomplishes these feats through an integration and combination of actions. Sapolsky (1994) provides one example of explanation from which we will draw to describe some of the difference between homeostasis and allostasis. Say we have a gasoline shortage in America. A homeostatic solution might be to build smaller engines for our cars. The allostatic approach would include: smaller car engines, tax rebates for car pooling, encouraging checking and maintaining proper tire pressure.

The term homeostasis has served us well for almost 80 years to describe changes on a small scale normally restricted to individual set points. Allostasis refers to an overall centrally mediated, orchestration aimed at maintaining viability, while adjusting numerous set points in preparation for or in response to a threatening situation. Sterling (2004) stated: “All scientific models eventually encounter new facts that do not fit, and this is now the case for homeostasis.” Allostasis describes an additional regulatory process of reestablishing homeostasis of the internal milieu through a physiologic change manifested to meet a real or perceived demand.

Allostasis and allodynamic regulation describe phasic arousal in field polygraphy where differential arousal is most likely associated with an emotional or motivational impetus. The short-term phasic arousal is well-described as an allostatic state. Allodynamic regulation describes the centrally mediated, integrative marshalling of bodily systems and resources to the perceived emotionality of the test questions.

The ever expanding science of psychophysiology has embraced this terminology and uses it when describing complex psychophysiologic interactions (Berntson & Cacioppo, 2007) and the science of polygraphy will benefit by the acceptance of sister disciplines. This is more likely to happen if we share a common language, one that adheres to the spirit of parsimony. If the polygraph profession is serious in its pursuit of general acceptance, it must be prepared to expand and embrace the common language and concepts of other sciences. Incorporating the concept of allostasis and allodynamic regulation into our profession is an important step in that direction.

References

Aggleton, J.P., and Mishkin, M. (1986). The amygdala: Sensory gateway to the emotions. In R. Plutchik and H. Kellerman (Eds.). *Emotion: Theory research and experience. Vol 3: Biological Foundations of Emotion*, 281-296. Orlando FL: Academic Press.

Bernard, C. (1878-1879). *Leçons sur les Phénomènes de la Vie Communs aux Animaux et aux Végétaux, 2 Vols*, Paris: Baillière. (Lectures on the phenomena of life common to animals and plants, 1974, translated by H.E. Hoff, R. Guillemin, and L. Guillemin. Springfield, Illinois: Thomas).

Berntson, G.G., and Cacioppo, J.T. (2007). Integrative Physiology: Homeostasis, Allostasis, and the Orchestration of Systemic Physiology, in Cacioppo, J.T., Tassinary, L.G., and Berntson, G.G. (Eds.). *Handbook of Psychophysiology, 3rd edition* (pp. 433-449). New York, NY: Cambridge University Press.

Bevan, A.T., Honour A.J., and Stott, F.H. (1969). Direct arterial pressure recordings in unrestricted man. *Clinical Science,* 36(2), 329-344.

Bolles, R.C. and Fanselow, M.S., (1980). A perceptual-defensive-recuperative model for fear and pain. *Behavioral and Brain Sciences*, 3, 291-301.

Cannon, W.B. (1928). The mechanisms of emotional disturbance of bodily functions. *New England Journal of Medicine*, 198, 877-884.

Cannon, W.B. (1932). *The Wisdom of the Body*. New York: Norton Press.

Davis, R.C. (1961). Physiological responses as a means of evaluating information. In *Manipulation of Human Behavior*, A. Biderman and H. Zimmer, (Eds.) New York: Wiley.

Department of Defense Polygraph Institute. (1994). *Anatomy and Physiology for the Forensic Psychophysiologist*. Ft. McClellan, AL, Department of Defense Polygraph Institute.

Dworkin B.R. (1993). *Learning and physiological regulation*. Chicago, Ill: University of Chicago Press.

Frijda, N.H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Gray, J.A. and Mc Naughton, N. (2003). *The Neuropsychology of Anxiety*, Oxford, Oxford University Press.

Gustafson, L.A. and Orne, M.T.(1963). Effects of heightened motivation on the detection of deception. *Journal of Applied Psychology*, 47(6), 408-411.

Heimer, L., Van Hoosen, G.W., Trimble, M., and Zahm, D.S. (2008). *Anatomy of Neuropsychiatry- The New Anatomy of the Basil Forebrain and Its Implications for Neuropsychiatric Illness*, Burlington, MA:Academic Press.

Janig, W. (2006). *The Integrative Action of the Autonomic Nervous System: Neurobiology of Homeostasis*. New York: Cambridge University Press.

Le Doux, J. E. (1996). *The Emotional Brain*. New York: Simon and Schuster.

Le Doux, J.E. (2002). *Synaptic Self- How Our Brains Become Who We Are*. Middlesex, England: Penguin Books Ltd.

McEwen, B.S. and Wingfield JC. (2003). The concept of allostasis in biology and biomedicine. *Hormones and Behavior,* 43:2-15.

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

Rosen, J.B., and Schulkin, J. (1998). From normal fear to pathological anxiety. *Psychology. Review*, 105, 325-350.

Sapolsky, R.M. (1994). *Why Zebras Don't Get Ulcers*, New York, NY: Owls Books, Henry Holt Company.

Schulkin, J. (2003). *Rethinking Homeostasis, Allostatic Regulation in Physiology*. Cambridge, MA: The MIT Press.

Sterling, P. (2004). Principles of allostasis: optimal design, predictive regulation, pathophysiology and rational therapeutics. In Schulkin, J. (Eds.). *Allostasis, Homeostasis, and the Costs of Adaptation*. Cambridge, MA: The Cambridge University Press.

Sterling, P., and Eyer, J. (1988). Allostasis: a new paradigm to explain arousal pathology. In: Fisher, S., Reason, J. (Eds.) *Handbook of Life Stress, Cognition and Health.* New York, NY: J. Wiley and Sons.

Walker, D.L., Toufexix, D.J., and Davis, M. (2003). Role of the bed nucleus of the stria terminalis versus amygdala in fear, stress, and anxiety. *European Journal of Pharmacology*, 463, 199-216.

Differing Perspectives but Shared Gratitude to a Person Instrumental in Advancing Polygraphy: David T. Lykken

Frank Horvath and Jamie McCloughan

In 1959 David T. Lykken received a proposal for research on polygraph testing from two of his graduate students at the University of Minnesota. At that time Lykken knew nothing about polygraph testing. He began a review of the literature to determine if there was a way to carry out the proposed research. The result was a combination of a word association and recognition format that he called the Guilty Knowledge Test (GKT). He carried out a study on the GKT and published the findings in "The GSR in the detection of guilt." The following year he published a second article in which he explored the resistance of the GKT to efforts to "beat the test." Thus was the genesis of the favored approach to "lie detection" amongst many in the scientific community. And, from that beginning one of the most vocal, prolific and recognized critics of the Comparison Question Technique (CQT) was born.

Late in 2006 David Lykken passed away, succumbing to what was, as he saw it, an overwhelmingly debilitating illness. While most examiners are likely to recognize Lykken by name and perhaps to know something of his position on "lie detection" there are other things examiners might also be interested in. Because both of us had personal experiences with Lykken we wish here to describe a few of these. We hope that sharing our experiences will help clarify Lykken's contributions to "lie detection" and at the same time make clear that in spite of our somewhat different observation points, we share the view that in the long run his influence will be beneficial.

There is no doubt that Lykken was a widely respected, extremely talented and influential psychologist and researcher. Given that, it is striking to us that some aspects of his position on "lie detection," particularly the CQT, were so misinformed. This is even more puzzling when one considers that he took time to visit at least one field examiner before he set out a "final" position in his written publications. For example, when he first started to write on and formulate a position on the CQT, he believed that "control questions" (now comparison questions, abbreviated as "CQs") were answered with a "yes," a "truthful" response. It was inexplicable to him that field examiners could honestly believe that such questions could provide useful comparison responses with which to evaluate physiological responses to relevant questions. When this error was pointed out to him, that CQs are usually answered with "no" (a probable "lie"), Lykken reported that this was even more inexplicable. Whether answered with a "no" (a "lie") or a "yes" (a non-lie) he seemed to think the idea of "control questions" was not just implausible but impossible. Importantly he didn't do any original research on the topic, though he did carry out one "real-life" examination in which he used a process similar to a CQ procedure in order to defend a person he thought had been wrongly accused.

Lykken's refusal to carry out any research on the CQ procedure was consistent with his position on some laboratory assessments. He dismissed all CQ laboratory research, believing that the CQs in that environment couldn't possibly work the way they do in real-life. In an actual investigation all suspects, truthful and deceptive, can readily identify the relevant and comparison questions; it was impossible to believe that truthful persons would respond more to the latter category of question than the former, given that the relevant question responses put them in jeopardy. On the other hand, Lykken (and others) saw no inconsistency in relying on laboratory assessments of the GKT and dismissing findings on the CQT in the same environment. This is still the position maintained by others, influenced strongly, no doubt, by Lykken.

Regardless of Lykken's stated position and his arguments against the CQT, it is possible, maybe likely, that he resisted the CQT not solely on conceptual or scientific grounds. The enmity between him and some proponents of the CQT in the "polygraph" field was not disguised. His position, in some respects may have reflected this personal side as much as it showed a professional view, something not uncommon in the controversy about the CQT. We believe it is likely that even if the accuracy of CQ testing were 100%, and proponents and skeptics alike agreed on that statistic, the issues of "civil liberties," the integrity of the justice system and "national security" as they relate to polygraph testing would remain. Those issues are more directly related to the way in which the CQT and polygraph testing is used than to the validity of the testing. Lykken's failure to address that fact and to consider the merits of CQT reveals only that he was opposed to CQ polygraph testing in spite of not solely because of the available evidence. He was, we believe, strongly supportive of "civil liberties" in a way that many in the polygraph examiner community would probably find disagreeable. In other words, ethical issues associated with the CQT were of special concern to Lykken, even though he was quite intrigued with the ability of CQT to perform well in certain instances.

Both of us corresponded with Lykken for a number of years. For one of us (FH) the correspondence was an exchange by mail in which we agreed to share our thoughts about polygraph testing. We agreed at the outset that our correspondence would be in confidence so that our views could be shared without fear that our honestly held convictions would not be abused.

At one point during that exchange, a state legislature was considering a bill to prohibit polygraph testing in employment situations. Although I was invited to attend, I was committed elsewhere and unable to be present at the hearings. Lykken, however, did attend. It is my understanding that the proponents of the legislation secreted Lykken, with his permission, in a back room until all other persons had testified. This being accomplished, Lykken was then called upon to present his arguments in favor of the legislation, without further rebuttal. During that testimony (a tape recording was made available to me) Lykken made a number of comments about persons in the polygraph field and about issues which were, in my judgment, extremely demeaning, unnecessary and without foundation. I wrote him about this and asked him to be considerate enough to, at the least, write to the legislators to clarify what had been said. He refused. Our correspondence stopped at that point, though we did talk briefly after that at a scientific meeting.

At times Lykken would make use of "evidence" in support of his position on the CQT which I am confident he would not do in other matters he wrote about. For instance, Lykken frequently referred to the case of Floyd "Buz" Faye as an example of how simple it was to "beat" the CQT. Faye provided Lykken with a testimonial stating that he had taught prisoners incarcerated with him how to "beat the test" based on Lykken's advice. Almost all of them were successful. While it is possible that Faye had been wrongly convicted and incarcerated, it is not likely, based on everything I read and learned about his situation, that what he claimed was true. Even if it were, however, there was no documented support for his assertions. Lykken, though, presented Faye's testimonial as if it were confirmed. I don't know if Faye was wrongly incarcerated but I do know about his polygraph examinations, as I was asked to review them. But, neither I nor any other examiner I am aware of ever reviewed those examination results which Faye maintained were evidence of the effectiveness of his advice to his fellow inmates. To my knowledge Lykken never reviewed those cases either.

Surprisingly, Lykken seemed to be unaware of field practitioners' historical development and use of testing procedures that were similar to the GKT. A clear example is what practitioners refer to as the Peak of Tension test. The POT is most certainly one of the procedures that would be included with the GKT in the larger family of "Information Recognition Tests" even though it does not have a scientific grounding in the way that the GKT does. Yet, Lykken's general lack of awareness of the POT and its prominence in the field literature suggest that he either ignored it or did not wish to take time to examine the literature.

I am not writing these words to disparage David Lykken. Rather, my point is to emphasize that the issues that separated him from some in the polygraph/research community were not always based on scientific premises. He was as passionate on his side as others are on the other side, if there are only two sides here. When it came to the GKT Lykken would go to great lengths to be helpful, constructive and supportive. When it came to the CQT Lykken's view was, we think, quite stubborn and fixed.

In 2002, one of us (JM) became interested in the use of the GKT in specific issue testing, something that was done on a daily basis as a polygraph examiner for a large police agency. I knew nothing of the GKT; it was neither discussed nor taught in the training school I attended. I purchased a used copy of Lykken's 1981 book "A Tremor in The Blood." After I completed the book, I sought out other information and contacted Lykken directly at his email address at the University of Minnesota. Surprisingly, he promptly replied to my inquiry and we began an exchange that would span almost four years.

Although I anticipated a lack of interest from Lykken, that was never the case. He was always prompt in his replies to my questions and he was always the professor. He would methodically address the issues I asked about and he would correct my errors in a constructive manner, always based on the available research. When they were required to support his points, he provided me with copies of articles and citations to research that he thought would be helpful.

On several occasions I called Lykken at his home. When I did he was always willing to discuss issues with me in some depth. I found him to be open to suggestions based on my experience with the use of the GKT in the field. He did not like the change in nomenclature that was being used to refer to the GKT as the "Concealed Information Test." But he said he understood why it had occurred.

After I had used the GKT for some time, I decided to write a procedural manual to guide me and others interested in the field implementation of the GKT. Lykken was instrumental in helping me to develop this manual. He reviewed the draft manual several times and provided comments on what he thought was acceptable and what required revision. His comments were quite detailed and in one case required me to revisit aspects of statistical analysis I thought would never apply in my law enforcement career. Lykken took particular interest in some of the terminology we used in the manual, such as the use of the word "control" for the incorrect items and "target" to refer to all of the items in a GKT, "key" and "controls". Remaining consistent with the terminology of a test, "key" still, as it did originally, referred to the correct alternative item.

Over the years of my use of the GKT, I would update Lykken on my progress or on the obstacles I encountered. Always, he had words of encouragement in return and he frequently voiced optimism that American polygraph examiners would one day surpass the Japanese in using the GKT. "I think I probably told you that the Japanese police use the GKT almost exclusively now, about 5,000 cases per year. Maybe we can catch up to them at last." (08/13/2004 email from David T. Lykken)

My last correspondence was sent to Lykken on September 13, 2006. His reply came the next day; it included a proposal for research and his ever present words of encouragement. At that time I was preparing a presentation on the GKT so I decided to wait until I returned to reply to Lykken. Sadly, when I got home I learned that Lykken had passed away the day after his last message to me.

I hope to continue to pursue the use of the GKT in the field and to help others who also wish to do so. In doing so the David Lykken I came to know, admire, and respect will remain affiliated with the field of polygraph testing.

Although we have shared our somewhat different experiences with David Lykken, each of us, in his own way, knows that the field owes a great debt of gratitude to this man. Aside from promoting the use of the GKT, there is little doubt that Lykken's influence and his prolific criticism of the field generally have led to a sense of urgency for more and better research in "lie detection" than would otherwise have been evident. Together we leave the reader with the last words of encouragement Lykken left to one of us: "Many thanks for the good news, and keep pitching!" (09/14/2006 email from David T. Lykken to JM). We, in our own ways, intend to do so.

Unsafe at any altitude:
Failed terrorism investigations, scapegoating 9/11, and the shocking truth about aviation security today.

Review by Frank Horvath

Unsafe at any altitude: Failed terrorism investigations, scapegoating 9/11, and the shocking truth about aviation security today, by Susan Trento and Joseph Trento, Hanover, New Hampshire: Steerforth Press, 2006.

New members of the APA won't know this but those who attended the seminar banquet in 1984 will recall that a record was established that evening: The person who holds the record for the longest speech by an incoming President is Frank Argenbright. In that speech he talked about his involvement in the polygraph field, his dedication to it and his successes and failures in the commercial aspects of it. His life course at that point did indeed have a lot of ups and downs; one could tell, though, that he was not a person to be daunted by setbacks. He simply worked through them or around them; they always succumbed to his efforts.

In the Fall of 2001, I was asked to join an ad hoc APA committee charged with developing an effective plan to implement a polygraph screening program for airport employees. For two days over one weekend we confined ourselves to a hotel meeting room and worked hard and long; we were driven by a consensus that there was real promise and value in what we were doing. A draft proposal was produced; over the next several weeks it went through several iterations. The final proposal, as I recall, was to be forwarded to Governor Thomas Ridge, who at the time was the Director of Homeland Security. I don't know if the proposal got to its destination or if anyone in a position of authority ever was made aware of it, much less read it. Too bad.

You might be asking yourself at this point what is the connection between my first paragraph and the second. How does Frank Argenbright relate to the APA's interest in promoting airport security? If you wish to fully understand the connection, I urge you to read this book. In fact, if you're a polygraph examiner, especially a private examiner, you should read this book. To my knowledge, Frank Argenbright was and still is the most commercially successful person ever involved in our field. But, like his life course years ago, detailed in his APA speech, his struggles have continued to be challenging, but not insurmountable for one with Frank's character. He is an unusually successful person.

True story: A number of years ago, as I tuned my television to a national news program, I happened, by coincidence, to catch an interview of a person wearing a paper bag over his head. He was being interviewed about a large charitable donation. I thought it was an odd scene and an odd way to achieve anonymity and it caught my attention. Also though I thought I recognized the person's voice: Frank Argenbright. As it turns out, it was him. That's one side of Frank that has not gotten much attention. Another side has; that, essentially, is the focus of this book.

Here's the other side. Frank Argenbright came from a modest background. He had a college degree (My observations suggest he struggled to get that, as other interests were more important to him at the time.) and a sincere interest in developing a successful polygraph business. He did that. His name became associated with a major training facility and a thriving commercial firm in Atlanta. In the course of that effort, however, he also "built a billion-dollar company by marshaling thousands (into a) highly motivated workforce that could provide low-cost security for the airlines. His formula was so successful that in just over twenty years his company became the largest aviation security firm in the United States, with 40 percent of the market….(he) took particular pride in the fact that his screeners had never been responsible for a serious security incident by allowing weapons through. Argenbright's methods were so successful in the United States that he was invited to revamp screening in Europe after the Pan Am 103 bombing over Lockerbie, Scotland. By 1999 he helmed the largest aviation security company in the world."

A pretty impressive picture, yes? All of that changed after September 11, 2001. A combination of factors led to disaster. Years before that incident, Argenbright set a personal goal of seeing his firm become a billion dollar company. He couldn't do that alone with private funds. He went public and his company, AHL became a billion dollar firm. Argenbright Security, however, with which his name was overtly affiliated, had been sold to a foreign firm, Securicor. Those were both on the "up" side for Argenbright; that didn't last long. Through the summer of 2001, during the dot.com bust, "there was nothing Frank Argenbright could do to keep AHL's (Argenbright's new company.) stock from falling. The economy was in a recession, and now the corporation had sold its core business (Argenbright Security-sold to Securicor)."

When 9/11 occurred Argenbright was in his boardroom negotiating a business arrangement for AHL. The receptionist rushed into the room and reported the news about a plane crashing into the World Trade Center. Argenbright learned later that day that the hijackers had passed through Argenbright Security checkpoints… and…. "it was beginning to look like airport screeners were not at fault---and all I could think of was thank God." This was the beginning of a long, steep "down" side for Frank Argenbright personally and professionally.

At the time of 9/11 Argenbright was not the owner of the company which bore his name, nor was he permitted to speak publicly for the company in response to 9/11. He and his family (and his "other" business) could do nothing but bear the tremendous pressure and demeaning response to the ostensible personal involvement of Frank Argenbright in 9/11. How they did this and what made it all so impossibly brutal is part of the story told in this book. Argenbright recalls, "I could not conceive that anyone would deliberately set out to divert attention from 9/11 by ruining my good name and the thousands of screeners who had worked very hard to make sure the passengers were safe. It just never occurred to me." But it happened.

The other focal point of this book should be of interest to all who are concerned about airport security. It starts with what Securicor's management -- Argenbright Security -- faced in the U.S. after 9/11. At the time it "was the biggest of the airport security companies in the United States, with ten thousand screeners at thirty-eight airports. Now the company was in serious danger because of a series of congressional proposals to simply eliminate private screeners." This was indeed the outcome.

"In the days after 9/11, lots of money and power were at stake, and forces were on the move in Washington. These competing forces had very little to do with either protecting the United States or mourning three thousand dead Americans. Instead the airlines, huge labor unions, federal agencies, Republicans, and Democrats all came together in a series of tawdry bipartisan political moves. What united these disparate forces was the opportunity to blame Frank Argenbright and the private airport screeners for 9/11. In the end, despite President Bush's devotion to the free market and private enterprise, the White House went along with a massive new federal bureaucracy, the Transportation Security Administration."

The TSA was put into operation in 2002. Since that time "weapons have been found behind the screening lines; bombs of identical design have been found at Seattle's SeaTac Airport and other hubs around the country. Not only was the nation's largest private aviation security firm, Argenbright Security, blamed falsely for allowing 9/11 to happen, but its experienced security screeners were unceremoniously dumped. In their stead, the government has hired thousands of convicted felons as security screeners; the theft of passenger property is commonplace. The worst news: The highly paid and well-informed forty-five-thousand-person-strong TSA screening force is much worse at detecting threats — bombs, explosives, and guns — than the private screeners they replaced. The TSA tries to keep the scores secret, but we learned that TSA screeners detect only about half the dangerous articles sent through airport security in tests. The private screeners routinely had an 80 to 95 percent detection rate. And not only are many of the TSA screeners incompetent, but many also take advantage of the federal system."

Perhaps of even greater interest to many in the APA is what is reported in this book regarding airport screening, something that could be and should be addressed more directly. It is made clear in this volume "that six hundred thousand employees had access to the back of America's airports. There was a deep, underlying fear among security people who understood how starkly the elaborate, sometimes arbitrary security required of passengers contrasted with the haphazard system for airline and airport employees and contractors." "The last group of people to have access to the plane before a flight… are the cleaning and service crews. What makes international air travel susceptible to terrorism is people who have access to the aircraft can be easily recruited either for ideological or financial reasons." An example of this was reported in the U.K.

"In recent years TSA has been so short-staffed it has diverted funds for improving liquid explosives detection technology to screener training. While tens of thousands of passengers wait for hours in airports and are forced to throw away all liquids and cosmetics for fear terrorists will bring aboard the ingredients for small bombs, the real danger in the British plot has been all but ignored by the media. The reason the British kept the security alert at a critical level several days after the plot was prematurely disclosed was that one of the British nationals arrested was a Heathrow airport security worker with an all-access pass to the huge facility. Amin Asmin Tariq was about to leave for work at Heathrow when British authorities arrested him at his home. One of the twenty-five people arrested in Britain in connection with the terrorist plot against British and American air carriers, Tariq was originally hired by GS4 (formerly Securicor — the company that bought Argenbright Security almost a year before 9/11) as an "ancillary security employee."

Aside from these examples, this book details the many problems yet to be faced in the airport security industry. Whether or not TSA or any federal bureaucracy will ever be as cost-effective as was the private sector, such as what Frank Argenbright built from a small bank account but a very powerful and very positive can-do attitude, remains to be seen.

Frank Argenbright is no longer actively and directly involved in the community of polygraph examiners. But, if all examiners gave a small bit of themselves to the field emulating Argenbright's enthusiasm for and dedication to success, the history of Polygraphy fifty years in the future will be much different than what is on the horizon now. I encourage you to read this book.

Epilogue:

The week after I finished writing this review an article appeared in *USA Today* entitled "TSA Inspections Follow Arrest of 2 Airport Workers." (Tuesday, March 13, 2007, p. 2A). The first paragraph stated: "The Transportation Security Administration began intense random inspections of airport workers after the arrest of two Orlando-based airline employees who allegedly carried 14 guns onto an airplane." One of the employees was said to have snuck "weapons and drugs onto an Orlando-to-Puerto Rico flight….[he] used his airline ID to board the flight carrying a duffel bag with 13 handguns, an assault rifle and 8 pounds of marijuana." The other employee was charged with helping the first person.

After this incident the Orlando airport immediately assigned more police to patrol employee-access doors. A spokesman for the House Homeland Security Committee said, "the added security is 'welcome but far from applause-worthy' because the TSA hasn't required screening of airport workers." The TSA recently required more airport workers to pass background checks. Is that going to help? A lot? The next time you stand in line in order to remove your belt and shoes, put your computer and personal supplies on a conveyor belt, have your identification checked and double-checked, and then walk through a magnetometer to gain access to an upcoming flight, remember, please, that a past-President of the APA, who also happened to grow a private security business into one of the largest in the world, had a better idea about airline security.