A Replication and Validation Study on an Empirically Based Manual Scoring System[1]

Ben Blalock, Barry Cushman & Raymond Nelson

Abstract

This is a replication of a study validating the hand scoring system for comparison question polygraph examinations proposed by Nelson, Krapohl and Handler (2008). Nine polygraph examiner trainees at an American Polygraph Association accredited polygraph school used an empirically based three-position manual scoring system involving three evaluative criteria and a reduced set of basic rules to evaluate 100 confirmed event-specific single-issue criminal investigation polygraph examinations from the Department of Defense Polygraph Institute confirmed case archive. Average decision accuracy for the inexperienced examiners was 88% with 13.1% inconclusives. Sensitivity and specificity levels achieved by the trainees did not differ significantly, suggesting they achieved balanced accuracy characteristics using the empirically based scoring system. All nine of the inexperienced examiners scored the sample cases with sufficient accuracy to meet the accuracy requirements specified by the Marin protocol (Krapohl, 2005; Marin, 2000). Results from this study parallel the results reported in the previous experiment and support the validity of an empirically based three-position manual scoring method.

Introduction

Nelson, Handler and Krapohl (2008) described a simplified empirically based manual scoring experiment and provided initial evidence that inexperienced polygraph examiners using a simple and empirically based system of test data analysis (hereafter the "Empirical Scoring System" or "ESS") appear to be capable of blind-scoring polygraph examination data with decision accuracy, inconclusive and interrater reliability rates that are equivalent to those of experienced scorers using existing, more complex test data analysis (TDA) methods (e.g., Krapohl & Cushman, 2006). The ESS method of TDA requires analysis of only the three simple and robust Kircher features, which have been the subject of numerous validation studies and which form the foundation of many computer-scoring methods (Dutton, 2000; Harris, Horner & McQuarrie, 2000; Honts & Driscoll, 1987, 1988; Kircher, Kristjansson, Gardner & Webb, 2005; Kircher & Raskin, 1988; Krapohl, 2002; Krapohl & McManus, 1999; MacLaren & Krapohl, 2003; Raskin, Kircher, Honts & Horowitz, 1988; Nelson et al., 2008), plus a fourth validated feature not currently utilized by computer scoring algorithms: temporary increase in respiration baseline (Bell et al., 1999).

Current measurement and statistically based approaches to TDA differ from older TDA methods, for which the emphasis is on pattern identification and pattern recognition (see Matte, 1996; Swinford, 1999). Because there are a greater number of features and more complicated criteria for determining a positive or negative score, it follows that they are more difficult to learn, remain somewhat more subjective, and can therefore be expected to ultimately produce weaker interrater agreement than measurement based systems. Although the present experiments involve a simple and empirically based manual scoring method that employs only those features for which there is convergent empirical validity in the published literature, the emphasis is not placed on

[1]The authors wish to express appreciation to Mark Handler, Charles Slupski, and Paul Kordonski for their comments and suggestions. The views expressed in this article are solely those of the authors, and do not necessarily represent those of the Lee County (Florida) Sheriff's Office, Portland (Maine) Police Department, Lafayette Instrument Company, or the APA. Comments or confirmed chart data should be emailed to blalock@InternationalPolygraph.com.

discrete or exact measurements, but instead depend on the simple and robust idea of visually comparing the relative magnitude of responses using a procedure known to polygraph field examiners as the *bigger-is-better* rule (Capps & Ansley, 1992a; Department of Defense Research Staff, 2006; Harwell, 2000; Krapohl, 1998; Van Herk, 1990).

Numerical scores are assigned using the three-position scale. Reactions of equal magnitude result in a score of zero while a visibly discernable difference in the magnitudes of the reactions being compared result in a score of plus or minus one, thus the rule "bigger is better." Ignoring the ratio of magnitude of the difference between two reactions reduces the subjectivity involved in trying to ascertain if a given reaction is "dramatic" enough, for example, and therefore deserving a score of a plus or minus two or three (Matte, 1996). The Empirical Scoring System, unlike the traditional 3-position scoring system, is designed to capitalize on the discriminating power of the EDA data over the other components (Capps & Ansley, 1992b; Kircher & Raskin, 1988; Kircher et al., 2005; Krapohl & McManus, 1999; Olsen, Harris & Chiu, 1994; Raskin et al., 1988). To do so, the EDA score is weighted more heavily than the other components. The weighting is achieved by assigning a score of a zero or a plus or minus two rather than a zero or a plus or minus one. Whereas a total question (per chart) score of four is possible, the EDA can attribute 50% of the score.

While computer scoring algorithms can execute complex evaluation and decision procedures with perfect reliability, it is axiomatic that increased complexity in human decision making can be expected to result in decreased reliability and decreased procedural adherence. Nelson et al. (2008) noted that computer scoring algorithms commonly evaluate simpler physiological features than those considered by human examiners. *The 2006 Federal Polygraph Examiner Handbook* (Department of Defense Research Staff, 2006) describes a reduced physiological feature set, compared to earlier scoring systems, that closely resembles that developed at the University of Utah (Bell et al., 1999) and that published by ASTM International (ASTM, 2009). Over time, it can be expected that further simplification and other improvement in polygraph test data analysis will lead to improved reliability and better acceptance by courts and other fields of science. The present study is a replication of Nelson et al. (2008), Experiment 4.

Method

Subjects

Nine polygraph students in their eighth week of polygraph training at an American Polygraph Association accredited polygraph school used the Empirical Scoring System outlined in Nelson, Krapohl, and Handler (2008) to evaluate 100 confirmed polygraph examinations. All examinations were from the DoDPI confirmed case archive; all were single-issue tests conducted as part of criminal investigations; and ground truth was verified independently of polygraph test outcomes. All nine participants were experienced law enforcement officers who were actively affiliated with various law enforcement agencies. All had received prior instruction in the current TDA procedures used by the Defense Academy for Credibility Assessment.

Procedure

Study participants were provided with 100 confirmed examination cases in PDF format. These 100 cases were taken from a confirmed database of examinations, with an equal number of truthful and deceptive cases (50 truthful, 50 deceptive) used. Confirmation of the cases came from either confession, DNA evidence, or from other irrefutable evidence. Each selected polygraph case was a single-issue examination, and was conducted employing the DoDPI (now DACA) Zone Comparison Technique (Krapohl & McManus, 1999).

They were provided instructions on how to employ the Empirical Scoring System, and they were instructed to score each case from the computer screen, without printing the charts or using mechanical or computerized measurement devices. Their scoring activities were limited to only assigning scores as described above: scores of plus, minus, or zero using the 3-position scale and the *bigger is better* rule (i.e., ignoring traditional scoring ratios). For the purpose of this experiment, participants were asked not

to render a decision regarding truthfulness or deception. Additionally, they were told to score each relevant question to its adjacent comparison question, or, where applicable, to the stronger of its two bracketing comparison questions for each component sensor. Simply stated, they were instructed to assign a point if they could visually discern a difference in magnitude between the segments being compared. They were further advised to not score reactions before the stimulus (i.e., reactions occurring too soon), or several seconds after the stimulus or answer (i.e., reactions likely not attributable to the question stimulus). No concrete instructions were provided regarding the required timeliness of the reactions, and the participants were asked only to refrain from scoring reactions that could be criticized as occurring to early or too late. (See Appendix A for a copy of the written instructions provided to the participants.)

Regarding the pneumograph waveforms, the participants were told to assign a score when one of three respiratory patterns are identified: 1) an increase in respiratory baseline following stimulus onset and containing three or more respiratory cycles before return to pre-stimulus baseline, 2) a suppression of respiratory amplitude of three or more respiratory cycles following stimulus onset, or 3) a slowing of respiration rate for three or more respiratory cycles from a consistent pre-stimulus level. Regarding the electrodermal data and cardiograph waveforms, participants were told that if they could discern a visible difference in the Y-axis or vertical amplitude of increase from baseline or lowest point following the stimulus onset (observed at the diastolic baseline in the cardiograph waveform), they should assign a score to the question with the greatest amplitude increase. Finally, they were told not to score data that were affected by movement artifacts, exaggerated or dampened response quality, or were of substantially unstable quality.

Data Analysis

All of the participants' scores were entered into a spreadsheet for evaluation purposes. As described earlier, all electrodermal scores were doubled to +/-2, regardless of the difference in response magnitude for relevant and comparison questions. Final decisions were made using two-stage scoring rules (Krapohl & Cushman, 2006; Senter, 2003; Senter & Dollins, 2002; Senter & Dollins, 2004) and statistically optimal cutscores (cutoffs) that were obtained from normative data and significance table (see Appendix B) reported by Nelson et al. (2008) using the OSS-3 decision thresholds of alpha ≤ .05 deceptive decisions, alpha ≤ .1 for truthful decisions, and a Bonferonni corrected alpha ≤ .017 for spot scores. *Two-stage scoring* involves evaluating scores assigned to individual question pairs (spots) in addition to total cutoffs with the aim of producing more accurate and less inconclusive decisions. The *spot score* is defined as when all of the scores assigned to a comparison and relevant question pair are added together from each presentation of the question pairs to create a total *spot score*. These values result in cutscores of +2 for truthful decisions, - 4 for deceptive decisions, and -7 for deceptive decisions based on spot scores alone.

Bootstrap mean and confidence intervals were calculated for decision accuracy, inconclusives, sensitivity, specificity, false negatives, and false positives. Using the statistically optimal cutscores, all nine participants in this study scored the data with sufficient accuracy to meet the requirements for Marin certification (ASTM, 2005). The bootstrap mean accuracy rate was 87.9% (95% CI = 81.4% to 94.5%), with an inconclusive rate of 13.2 % (95% CI = 6.7% to 19.8%). Bootstrap mean sensitivity to deception was .79 (95% CI = .68 to .90) and specificity 73 (95% CI =.61 to .86). The bootstrap mean false negative error rate was .11 (95% CI =.02 to .19), and the false positive error rate was also .11 (95% CI =.02 to .19). All of the observed mean scores were within the 95% confidence intervals reported in the Nelson et al. (2008) study. Kappa (inter-scorer agreement) for these nine participants was .56, which does not differ significantly from that reported by Nelson et al. (2008), and the bootstrap 95% confidence interval for the Kappa statistic was .48 to .63, suggesting good interrater agreement is achieved with the empirically based scoring system.

To further evaluate the role of pneumograph data in the empirically based scoring paradigm, additional analysis was

conducted on the pneumograph scores and waveforms. Using the automated measurements from the OSS-3 computer algorithm, along with a *bigger-is-better* decision scheme in which standardized logged values greater than zero are deemed NSR while standardized logged values lower than zero are deemed SR, the RLL method (measuring line length from question onset for ten seconds) gives 69% decision accuracy, with 75% for deceptive and 66% for truthful cases. After evaluating the participants' hand scores, in which pneumograph pattern recognition *alone* was utilized to diagnose truth or deception, the pneumograph hand scores, using values of greater than zero resulting in a conclusion of NSR and values of less than zero resulting in a conclusion of SR gives provided 55.1% overall decision accuracy with 55.3% for deceptive and 54.9% for truthful cases. Using a simple test of proportions, this difference was statistically significant different from chance ($p < .01$).

Discussion

This study was conducted as a replication of an earlier study by Nelson et al. (2008), and it provides support for the validity of the Empirical Scoring System. The Empirical Scoring System is based on simple and robust ideas that are well supported in published studies and have face validity. The inexperienced examiners (trainees) in this study scored polygraph charts at accuracy and reliability rates consistent with those of the experienced examiners reported by Krapohl and Cushman (2006), which should be of interest to trainers, field examiners and program managers. If it is reasonable to assume that field experience is valuable and contributes to increased skill and performance in test data analysis, then the performance of the inexperienced scorers might be attributable to an improved emphasis on empirically sound principles in their scoring method.

The Empirical Scoring System, based on the *bigger-is-better* rule, is not only straightforward to use, but is also easy to explain to polygraph examiners and non-examiners such as department administrators or adjudicators, and it offers promising potential for gaining increased understanding and increased credibility among consumers of polygraph test results. The principle of weighting the contribution of the electrodermal component more heavily than other components, which was previously described in the existing literature, is further supported by the present study as is the use of two-stage decision policies. Finally, the present study provides support for the hypothesis that automated measurement-based scoring approaches, utilizing RLL measurements of the pneumograph waveforms, offer the potential for additional increased accuracy over pattern recognition scoring methods.

Given the consistently high criterion validity and good interrater reliability of the Empirical Scoring System, we recommend its use among examiners and program administrators as an expedient method for interpreting the results of polygraph tests in field settings. An additional advantage of the ESS, compared to existing hand-scoring systems, is the existence of normative data that can be used to provide an expedient understanding of the level of statistical significance achieved by various decision cutscores (see Appendix B). In an era that emphasizes theoretically sound decision models, mathematically defensible results, and known methods for calculating the likelihood of an erroneous test result, all investigators involved in development and research of polygraph scoring systems should feel an obligation to publish normative data and significance tables for all polygraph scoring systems in present use. A limitation of the present study is that it applies only to Zone Comparison Tests using three investigation targets that describe a single known incident or known allegation. Questions will undoubtedly arise as to what to do in the event vasomotor activity or additional charts (or both) are collected. Unless and until future research suggests a need to alter cutscores, it is recommended examiners utilize the cutscores described herein for all such exams as is the current practice with the seven-position Utah Scoring System (Bell et al., 1999). Generalization of the Empirical Scoring System to multi-facet exams using the Modified General Question Technique, and the application of the presently available normative data and significance table to mixed issue screening exams will require further development.

References

ASTM (2005). E2324-04 Standard Guide for PDD Paired Testing. ASTM International.

ASTM (2009). E2229-09 Standard Practices for Interpretation of Psychophsysiological Detection of Deception (Polygraph) Data. ASTM International.

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

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

Capps, M. H. & Ansley, N. (1992b). Analysis of private industry polygraph charts by spot and chart control. *Polygraph*, 21, 132-142.

Department of Defense Research Staff (2006). *Federal Psychophysiological Detection of Deception Examiner Handbook*. available online: Retrieved 1-10-2007 from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf.

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

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

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

Honts, C. R. & Driscoll, L. N. (1987). An evaluation of the reliability and validity of rank order and standard numerical scoring of polygraph charts. *Polygraph*, 16, 241-257.

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, p. 39463.

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., Kristjansson, S. D., Gardner, M. K. & Webb, A. (2005). *Human and computer decision-making in the psychophysiological detection of deception*. : University of Utah. Final Report

Krapohl, D. (2002). The polygraph in personnel screening. In M. Kleiner (Ed.), *Handbook of Polygraph Testing* (). San Diego, CA: Academic Press.

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

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

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.

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.

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

Nelson, R., Krapohl, D. J. & Handler, M. (2008). Brute Force Comparison: A Monte Carlo Study of the Objective Scoring System version 3 (OSS-3) and Human Polygraph Scorers. *Polygraph* , 37, 185-215.

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

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

Senter, S. & Dollins, A. (2004). Comparison of Question Series and Decision Rules: A Replication. *Polygraph*, 33, 223-233.

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

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

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.

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, p. 28-47. Reprinted in *Polygraph*, 20(2), 70-79.

Appendix A

Empirical Scoring System -Scoring Instructions

Scoring

Assign values of +, - or 0 using the 3-position system and the *bigger is better principle*

- Score each relevant question to the stronger of bracketing comparison questions, for each component sensor
- Do not be concerned about traditional scoring ratios
- If you can visually (without mechanical or automated measurement) determine that one segment is larger than another, then you can assign a point

Physiological Signals

1. Respiratory Suppression
    - Decrease in respiration amplitude for three respiratory cycles, beginning after the stimulus onset
    - Decrease in respiration rate (slowing) for three respiratory cycles, beginning after the stimulus onset
    - Temporary increase in respiratory baseline for three respiratory cycles, beginning after the stimulus onset
2. Electrodermal amplitude of increase
3. Cardiograph amplitude of increase, measured at the diastolic baseline

Interpretation Rules

1. Score only timely reactions
    - Do not be concerned about traditional scoring periods
    - Do not score reactions that begin before the stimulus
    - Do not score reactions that begin long (several seconds) after the stimulus or answer
2. Score only normal interpretable data
    - Do not attempt to score data that are affected by movement artifacts
    - Do not attempt to score messy or unstable segments of data
    - Do not attempt to score data of unusual response quality (dampened or exaggerated)
3. Double all EDA scores to +/-2

Significance Table for ZCT Decision Cutscores (Nelson, Handler & Krapohl, 2008)

Distribution of Deceptive Scores			Distribution of Truthful Scores	
NSR Cutscore	Z-value (alpha)		SR Cutscore	Z-value (alpha)
-1	0.154		-11	0.004
0	0.127		-10	0.006
1	0.104		-9	**0.008**
2	***0.085***		-8	0.012
3	0.068		***-7***	***0.017***
4	0.053		-6	0.023
5	***0.042***		-5	0.032
6	0.033		***-4***	***0.042***
7	0.025		-3	0.056
8	0.019		-2	0.073
9	0.014		-1	0.093
10	0.010		0	0.118
11	0.007		1	0.146

Mean deceptive score = -9.63 (SD = 8.47)

Mean truthful score = 8.85 (SD = 7.46)