Credibility Assessment Using Bayesian Credible Intervals:

A Replication Study of Criterion Accuracy

Using the ESS-M and Event-Specific Polygraphs with Four Relevant Questions

Raymond Nelson[1]

Criterion accuracy was evaluated for event-specific polygraph exams with four relevant questions. The sample included n=15 innocent cases and n=15 guilty cases, selected from an archive of confirmed field cases that was compiled by the Department of Defense in 2002. All cases employed relevant questions that described both direct and indirect involvement in the crime under investigation. Physiological responses were extracted from recorded computer software designed to execute the scoring procedures described in the published literature. Numerical scores were assigned using the Empirical Scoring System. A multinomial likelihood function was used to calculate a statistical value for the numerical scores. The cases were classified as either deceptive, truthful or inconclusive using two-stage decision rules and a naïve-Bayes classifier for which the 95th percentile limit of the credible interval for the posterior odds of deception or truth-telling was calculated using the Clopper-Pearson method. They were classified as deceptive when the 95th percentile limit of the credible interval for the posterior odds of deception exceeded the prior odds of one to one. Similarly, the samples were classified as truthful when the 95th percentile limit of the posterior odds of truth-telling exceeded the prior odds. Results for two of the sample cases (7%) were inconclusive because the 95th percentile limit of the posterior odds did not exceed the prior odds. Correct classifications were made for 93% of the 28 cases where the posterior odds were statistically significant (where the 95th percentile limit of the credible interval exceeded the prior odds). Test sensitivity to deception was observed at .87, and test specificity to truth-telling was also observed at .87. These results are consistent with previously published descriptions of event-specific examinations with four relevant questions.

Previous publications have described the structure of event-specific examinations with four relevant questions, known to field examiners as variants of the MGQT format, including the AFMGQT version 1 and version 2 (Department of Defense, 2006a, 2006b), and the Utah four-question format (Handler & Nelson, 2008; 2009), known to some field examiners as the "Raskin technique" due to his role in the development of this approach. These examinations can consist of both primary relevant questions that describe an examinee's direct involvement in the issue under investigation, along with secondary relevant questions that attempt to describe an examinee's indirect involvement or level of involvement. These examinations are traditionally interpreted with an assumption of independent criterion variance. However, previous studies have not supported the validity of the independence hypothesis for these examination for-

1 Raymond Nelson is one of the developers of the Empirical Scoring System (ESS) and Objective Scoring System, version 3 computer scoring algorithm, and has published numerous studies on all aspects of the polygraph test. Mr. Nelson is a polygraph field examiner and psychotherapist with expertise in sexual offending, victimization, trauma and development in addition to other experience in testing, data analytics and statistics. Mr. Nelson serves as an expert witness in legal matters involving both polygraph and psychology/psychotherapy. Mr. Nelson is a past president, and currently elected member of the APA Board of Directors, and has helped with policy development at the state, local and national level. Mr. Nelson is a research specialist with Lafayette Instrument Company, which develops and markets polygraph technologies. The views an opinion expressed herein are those of the author and not the APA or LIC.

mats (Barland, Honts & Barger, 1989; Podlesny & Truslow, 1993).

In response to inquiry from field practitioners, Nelson, Handler, Oelrich and Cushman (2014) described the rationale for a more generalized usage of polygraph test formats for event-specific diagnostic exams. Nelson and Handler (2017) described a general procedural rationale for the selection of a test formats for screening and diagnostic exams, indicating that while statistical multiplicity may lead to reduced precision for screening with more relevant questions, precision of diagnostic exams may be increased through the use of more relevant questions and the acquisition of a greater volume of test data.

Raskin, Honts, Nelson & Handler (2015) reported the results of a Monte Carlo analysis of these examinations, and suggested that criterion accuracy with four relevant questions can equal or exceed that or other examination formats when these exams are evaluated with an assumption of non-independent criterion variance. The present study was designed as a replication, in an attempt to increase the available published information about four relevant question event-specific polygraphs, including: test sensitivity, specificity, false-negative and false-positive errors, and inconclusive results using confirmed field cases.

Data

Examination data for this study were obtained from an archive of confirmed field polygraph exams that was compiled at the Department of Defense in 2002. Cases consisted N=30 confirmed field polygraph examinations. There were n=15 confirmed innocent examinations of this type in the confirmed case archive, along with a random selection of n=15 matching confirmed deceptive cases. Investigation target issues included: theft/larceny, murder, sexual assault, aggravated assault, false statements/false swearing, arson, robbery, child abuse, fraud, and illegal drugs. Archival data indicate that all sample cases were confirmed by information other than examinee confession. All examinations consisted of four relevant questions presented in a sequence with other questions designed to elicit responses that can be compared with responses to the relevant test stimuli, in addition to other procedural questions. Relevant questions included a combination of questions about direct involvement and indirect involvement in the issue under investigation. An important characteristic of contemporary field polygraph test formats is that all relevant questions subject to numerical evaluation are presented subsequent to the presentation of at least one comparison stimuli. Another important characteristic is that responses to each of the relevant stimuli are evaluated using the comparison stimulus immediately preceding and immediately subsequent to the relevant stimuli depending on which comparison stimuli has elicited the greater change in physiological activity.

Analysis

Data for each case was exported to a structured ASCII text format, including time-series data for all recording sensors, along with event markers indicating the onset, end and verbal answer for all test stimuli, along with other annotations. Data were imported to the R statistical computing environment (R Core Team, 2017) for signal processing and feature extraction. Response features were those described in previous publications (Krapohl & McManus, 1999; Nelson, Krapohl & Handler, 2008). Those features include: amplitude of increase in EDA, amplitude of increase in blood pressure, and suppression or reduction of respiration activity. Numerical scores were assigned to each of the sensors for each stimulus presentation using the Empirical Scoring System (Nelson et.al., 2011).

Posterior odds of deception or truth-telling were calculated for each case using a multinomial likelihood function for ESS scores (ESS-M) and a naïve-Bayes classifier (Nelson, 2017). The 95th percentile one-tailed limit of the Bayesian credible interval was calculated using the Clopper-Pearson method. Classifications of deception or truth-telling were made using two-stage rules (TSR; Senter, 2003; Senter & Dollins, 2003). The TSR requires that cases would be classified as deceptive when the 95th percentile limit of the credible interval for the posterior odds of deception exceeded the prior odds of one to one using the grand total score. Similarly, cases would be classified as truthful when the 95th percentile limit of

the posterior odds of truth-telling exceeds the prior odds using the grand total score. When results are inconclusive using the grand total score, the TSR would permit a deceptive classification if the 95th percentile limit of the multiplicity-corrected posterior odds of deception for the lowest subtotal score has exceeded the prior odds. Cases would be unclassified, and therefore inconclusive, when 95th percentile limits of the grand total and lowest subtotal score have not exceeded the prior odds.

Results

The mean score for innocent cases was 13.3 (sd=10.0), and the mean score for guilty cases was 14.1 (sd=12.7). Results with the naive-Bayes classifier and ESS-M scores are shown in Table 1. Two cases were inconclusive, including one guilty and one innocent case. In addition one of the innocent cases was incorrectly classified as deceptive, and one of the guilty cases was incorrectly classified as truthful. Twenty-six of the cases were classified correctly. A detection efficiency coefficient (Kircher, Horowitz & Raskin, 1988) was calculated in order to provide a single statistical metric to encompass correct, incorrect, and inconclusive results with both guilty and innocent cases. The detection efficiency coefficient was .83.

Excluding inconclusive results, 93% of the decisions from the naive-Bayes ESS-M classifier were correct. Several metrics of classification accuracy were calculated, including test sensitivity to deception, specificity to truth-telling, false-negative and false-positive errors, inconclusive results, and unweighted criterion accuracy. Confidence intervals, shown in Table 1, were calculated for all metrics using a parametric bootstrap.

Table 1. Criterion accuracy of ESS-M scores of event-specific exams with four relevant questions.

Unweighted accuracy	.93 {.87 to .98}
Unweighted inconclusive	.07 {.02 to .12}
Sensitivity	.87. {77. to .95}
Specificity	.87. {77. to .95}
False negative	.07 {<.01 to .14}
False positives	.07 {<.01 to .14}
Guilty inconclusive	.07 {<.01 to .14}
Innocent inconclusive	.07 {<.01 to .14}

Results for two of the sample cases (7%) were inconclusive because the 95th percentile limit of the posterior odds did not exceed the prior odds. Correct classifications made for 93% of the 28 cases where the posterior odds were statistically significant (where the 95th percentile limit of the credible interval exceeded the prior odds). Test sensitivity to deception was observed at .87, and test specificity to truth-telling was also observed at .87. These results are consistent with previously published descriptions of event-specific examinations with four relevant questions. Incorrect classifications were made for two of the sample cases, including one innocent case (7%) and one guilty case (7%).

Discussion

This project was an attempt to replicate previous work on event-specific diagnostic polygraphs with four relevant questions. This project also replicates previous work involving the use of a multinomial referenced distribution and naive-Bayes classifier for ESS-M scores. Results from this study are consistent with other reported results involving event-specific polygraphs with four relevant questions evaluated with an assumption of non-independent criterion variance.

To further investigate the differences between the traditional approach to these examinations and results using an evidence-based statistical classifier, the detection efficiency coefficient and results were re-calculated using the subtotal-score rules (SSR) and traditional numerical cutscores. Traditional numerical cutscores for these examinations are +3 for truthful classification of the subtotal scores, and -3 for deceptive classification of the subtotal scores. The SSR requires that all subtotal scores exceed the traditional numerical cutscore in order to classify a case as truthful, while any deceptive subtotal score would result in a classification of the case as deceptive. The SSR does not permit both truthful and deceptive decision within a single exam. The detection efficiency coefficient using the SSR and traditional cutscores was .75. Of the 30 cases, 10 (33%) were inconclusive using the SSR and traditional cutscores, including 1 guilty case and 9 innocent cases. Correct classifications of deception and truth-telling were made for 14 of the guilty cases (93%) and 4 of the innocent cases (27%). The unweighted accuracy, excluding inconclusive results, was .83 using the TSR and traditional cutscores.

Results from this project suggest that decision accuracy could benefit substantially from a change from the traditional decision rules to others for which published evidence has found better performance.

Traditional approaches using the SSR for the interpretation of polygraph tests that use a combination of four primary and secondary relevant questions are known to produce accuracy rates that underperform compared to other well-known testing approaches. Interpretation of these exams using the evidence-based TSR and cutscores that are informed by sound statistical theory can produce classification accuracy rates that may equal or exceed that of other highly-regarded polygraph formats for event-specific diagnostic exams.

Like all projects, this project is not without limitations. Among the obvious limitations herein, is the small sample size. Additionally, incomplete information was available regarding the examinee demographics, and no information is available concerning how the sample cases came to be included in the confirmed case archive. These limitations notwithstanding, the present results support a recommendation for continued interest in the four relevant question event-specific format for field practitioners and researchers within the polygraph profession. In addition, these results support continued interest in the TSR and the ESS-M naive-Bayes model for statistical quantification and classification of polygraph test results.

References

Barland, G. H., Honts, C. R. & Barger, S.D. (1989). Studies of the accuracy of security screening polygraph examinations. DTIC AD Number A304654. Department of Defense Polygraph Institute.

Clopper, C. & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 26, 404–413.

Department of Defense (2006). Federal Psychophysiological Detection of Deception Examiner Handbook. Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02- 10-2006.pdf on 3-31-2007. Reprinted in Polygraph, 40(1), 2-66.

Department of Defense (2006). Psychophysiological Detection of Deception Analysis II -- Course #503. Test data analysis: DoDPI numerical evaluation scoring system. Available from the author. (Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007).

Handler, M. & Nelson, R. (2008). Utah approach to comparison question polygraph testing. European Polygraph, 2, 83-110.

Handler, M. & Nelson, R. (2009). Utah approach to comparison question polygraph testing. Polygraph, 38(1), 15-33.

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., Horowitz, S. W. & Raskin, D.C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. Law and Human Behavior, 12, 79-90.

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

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

Nelson, R. & Handler, M. (2017). Practical polygraph: how to select a polygraph test format. APA Magazine 2017, 50 (2), 72-81.

Nelson, R., Handler, M., Oelrich, M. & Cushman, B. (2014). APA Research Committee Report:proposed Usage for an Event-specific AFMGQT Test Format. Polygraph, 2014, 43(4), 155-167.

Nelson, R. Handler, M. Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B., & Oelrich, M. (2011). Using the Empirical Scoring System, Polygraph, 40(2).

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

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.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raskin, D. C., Honts, C., Nelson, R. & Handler, M. (2015). Monte Carlo estimates of the validity of four relevant question polygraph examinations. Polygraph, 44(1), 1-27.

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

Senter, S. M. & Dollins, A. B. (2003). New Decision Rule Development: Exploration of a two-stage approach. Report number DoDPI00-R-0001. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC. Reprinted in Polygraph, 37(2), 149-164.