Reducing Inconclusive Results: A Descriptive Analysis of Decision Rules, Weighted Electrodermal Scores and Multinomial Cut-scores

Raymond Nelson and Mark Handler[1]

Abstract

An archival sample of N=100 confirmed field polygraph exams was used to calculate descriptive statistics, including point estimates, to study the effect on inconclusive results and other metrics of test accuracy. Data were analyzed as a function of different decision rules, structural weighting of sensor scores and cut-scores. Decision rules included the federal-zone-rules, grand-total-rule, subtotal-score-rule, two-stage-rules. Numerical scores were obtained through an automated feature extraction algorithm. Results were evaluated with both unweighted three-position numerical scores and after weighting the integer scores for electrodermal responses. Results are shown for both traditional numerical cut-scores and also using cut-scores obtained from multinomial reference distributions for three-position comparison question test scores with both weighted and unweighted electrodermal scores. Use of different decision rules had less effect with the multinomial cut-scores than with traditional cut-scores. Weighted EDA scores produced an average 49% reduction of inconclusive results across all decision rules, and the combination of weighted EDA scores and multinomial cut-scores reduced the occurrence of inconclusive results by an average of 72%.

Introduction

Few things are more disappointing for field polygraph examiners and referring agents than an inconclusive[2] test results. There is little to compare with the sense of frustration when – after all the time and effort invested in preparing for the test, the pretest interview, target selection, question formulation, test data collection, and test data analysis – a test result is not statistically significant for deception or truth-telling. This sense of frustration is, at times, shared by some examinees – especially those who are innocent – who, having agreed to testing in hopes of producing test data that can serve as a basis of evidence to support professional conclusions about innocence or truth-telling, must necessarily be informed that their inconclusive test result provides no better information about truth or deception than was already available before
the test.

In years past, in the absence of a probabilistic view of polygraph test results, it may have been tempting to up-sell the capabilities of the polygraph as virtually infallible. During that era, an inconclusive result was at times regarded as an indication of an unskilled examiner. Behind this attitude or belief was likely a sincere desire among polygraph examiners to be of actual help to referring professionals, for whom an inconclusive test result offered little or no practical value. And virtually nobody wants to purchase the services of a probabilistic test for which we could offer little more than conjecture as to the strength of the conclusion. In the absence of an ability to realistically quantify the level of confidence

1 The authors are extremely grateful to Mr. Don Krapohl and Texas DPS Captain Matt Hicks (who functioned as the action editor for this project) for reviewing, commenting and editing earlier drafts of this manuscript.

2 To remain consistent with the terminology of the forensic sciences we have chosen to call indeterminate polygraph results “inconclusive” rather than the neologism “No Opinion.”

or margin of uncertainty for the test result, it may have been a matter of professional marketing to up-sell or over-sell the capabilities of the test, including unrealistic expectations for infallibility and the absence of inconclusive results.

It may be helpful to remember that in- conclusive results are not unique to polygraph testing; all forensic tests are burdened with a certain proportion of inconclusive outcomes. The stated reason for them depends upon the forensic discipline in which the test is conducted, but in the main it is because the signal/ pattern/trace/sample/marker/image is inadequate or contaminated. So too is it with the polygraph. Even under ideal conditions, with heroic effort and with perfect examinees, pre-test interviews, and charts there will always be some cases in which the data do not allow for a reliable decision. An important objective in all forensic disciplines is to minimize those occasions.

Objectively, if people believed the poly- graph to be infallible then there would be no great difficulty in accepting a test result at face value. However, most polygraph examiners, and other professionals that we are aware of, would be hesitant to accept a test result at face value without scrutinizing the test administration, test data, and analytic result. Also, no thinking person today will accept the notion that polygraph is infallible or deterministic. It is well established that all scientific test results rely on probabilities and probability theory to quantify important phenomena that cannot be subject to perfect deterministic observation or direct physical measurement. Polygraph technologies, methodologies, and standards have all evolved along with this understanding.

Presently, it is reasonably understood by most – polygraph professionals, referring agents, courts, legislators, scientists, members of the community, and the entertainment and news industries – that the polygraph test is, like other scientific tests, merely a statistical classifier intended to quantify a phenomenon that cannot be subject to perfect deterministic observation or direct physical measurement. All tests are fundamentally probabilistic. Along with any use of statistics and probability theory comes the potential for testing error. The practical purpose of test data analysis is often to achieve a categorical classification or categorical test result. But experts who possess a broader and more complete understanding of scientific test are aware that the actual purpose of test data analysis is to quantify – in some reproducible manner – the level of confidence or margin of uncertainty surrounding a test result. Any reasonable and intelligent use of statistical theory and statistical methods will include an acknowledgement of some potential for testing error, and some potential that test results may not achieve a required level of statistical significance. Despite this, inclusive test results, are still the bane of field polygraph examiners everywhere.

Reducing the proportion of inconclusive results can potentially increase the effectiveness of the test. Inconclusive results are inversely related to utility. Higher rates of inconclusive results correspond with a lower rate of utility or usefulness. One common strategy for reducing the occurrence of inconclusive test results is to increase the volume of available test data by repeating the series of relevant and comparison questions one or two additional times, beyond the minimum three repetitions. Another strategy for reducing the occurrence of inconclusive results is to include additional independent diagnostic information, such as from a vasomotor sensor, to the traditional array of respiration, cardio, and electrodermal sensors. There are also other, subtler, strategies to reduce the occurrence of inconclusive test results, and these can include target selection, question formulation, and interviewing skills.

In this paper we use an archival sample of confirmed criminal investigation polygraphs to show three additional evidence-based approaches to field practices that can reduce the occurrence of inconclusive test results while increasing the objectivity and statistical power of the test. The first among these involves *polygraph decision rules*. The second field practice area involves the *weighting of electrodermal* (EDA) scores when assigning Likert-type (1932) three-position scores to physiological response to polygraph questions. Thirdly, we show the effect of *replacing traditional polygraph cut-scores* with cut-scores derived from a multinomial distribution of polygraph scores based the analytic theory of the polygraph test.

Sample Data Method

procedures can be found in publications by the Department of Defense (2006a; 2006b) and is referred to as the Federal Zone Rule (FZR) for the remainder of this manuscript.

A sample of confirmed field polygraphs was obtained from an earlier study by Krapohl (2005). The sample data consisted of N=100 polygraph examinations that were selected randomly from an archive of confirmed field cases. All exams were conducted by U.S. law enforcement agencies using the Federal Zone Comparison (ZCT) test format (Light, 1999; Department of Defense, 2006a), for which the sequence of scored questions, including three relevant questions and three comparison questions, was repeated three times. Sample cases consisted of n=50 confirmed truthful cases and n=50 confirmed deceptive cases. Data for all cases consisted of recording sensors for changes in thoracic and abdominal respiration activity, electrodermal activity, and cardiovascular activity.

Table 1 shows is an extended calculation of the sample results reported by Krapohl (2005) using the Federal Zone Rule (FZR) and seven-position scores, including test sensitivity or true-positive (TP) rate and specificity or true-negative (TN) rate, false-positive (FP) and false-negative (FN) errors, inconclusive (INC) results, and the unweighted average of correct decisions and inconclusive results for confirmed deceptive and truthful cases. Table 1 also shows the positive-predictive-value (PPV; calculated as TP/(TP+FP) and negative-predictive-value (NPV; calculated as TN/(TN+FN). Also shown in Table 1 is the detection efficiency coefficient (DEC; Kircher, Horowitz & Raskin, 1988), calculated as the Pearson correlation of the case status [-1, 1] and test result [-1, 0, 1]. The DEC provides a single metric that encompasses correct classifications, errors and inconclusive results for both deceptive and truthful sample cases.

Krapohl (2005) reported results using "investigative rules," for which deceptive classifications were made if the grand total score equaled or exceeded a cut-score of -6, or any subtotal score equaled or exceeded -3. Truthful classifications were made only when the grand-total score equaled or exceeded +6 *and* all subtotal scores exceed +1. All other conditions were classified as inconclusive. These

Krapohl (2005) also reported results using "evidentiary rules," for which deceptive classifications were made if the grand total score equaled or exceeded -6 while truthful classifications were made if the grand total score equaled or exceeded +4. In cases where the grand total was in the range from -5 to +3 the subtotal scores were used to make deceptive classifications if any subtotal equaled or exceeded -3. All other conditions were classified as inconclusive. The process of first using the grand total score and subsequently using subsequently using subtotal scores if the grand total is inconclusive was first described by Senter and Dollins (2003) and is referred to as the two-stage-rule (TSR) for the remainder of this manuscript.

Table 1. Sample results (n=100) reported by Krapohl (2005) for seven-position scores.

	Investigative Rules / FZR	Evidentiary Rules / TSR
Sensitivity (deception)	.78	.81
Specificity (truth-telling)	.62	.80
False-negative Errors	.07	.10
False-positive Errors	.09	.09
Inconclusive-guilty	.15	.09
Inconclusive-innocent	.29	.11
Unweighted Accuracy	.90	.90
Unweighted Inconclusives	.22	.10
Positive Predictive Value	.90	.90
Negative Predictive Value	.90	.89
Detection Efficiency Coefficient	.67	.74

Results from Krapohl (2005), shown in Table 1, results are consistent with other studies using the FZR and showed that in- conclusive results are loaded for innocent persons. Krapohl showed that use of the seven-position scores with the TSR and evidentiary cut-scores resulted in similar classification accuracy, though with improved test specificity and a reduction of the inconclusive rate by 55%. Importantly, Krapohl used different cut-scores for the FZR and TSR, and it remains unknown what portion of the observed difference can be attributed to decision rules and/or to the different cutscores. The present analysis is an attempt to provide more information about observed differences in test accuracy as a function of inconclusive rates for decision rules, cut-scores and other factors.

Feature Extraction and Data Reduction

Sample case were scored using a three-position numerical scoring method (Bradley & Janisse, 1981; Department of Defense, 2006b; van Herk, 1991). For each iteration of each relevant question and for each recording sensor, three-position numerical scores were assigned via an automated feature extraction algorithm that was developed using the R statistical computing language (R Core Team, 2018). The three-position scoring method is Likert-type coding system, based on the analytic theory of the polygraph test (Nelson, 2015). Scores of +1, 0 and -1 were assigned to relevant and comparison question pairs – referred to by field polygraph examiners as *analysis subtotals or "spots".* Negative scores were indicative of greater changes in physio- logical activity in response to relevant questions whereas positive scores were indicative of greater changes in physiological activity in response to comparison questions. Scores of 0 were assigned when there was little or no difference in response to relevant and comparison questions.

The automated algorithm was designed to extract information about the relative amplitude of increase in electrodermal activity, relative increase in blood pressure, and relative suppression or reduction of respiration activity. These responses have been shown to be correlated with differences in physiological response to relevant and comparison polygraph questions under the analytic theory of the comparison question polygraph test (Kircher & Raskin, 1988; Kubis, 1962; Summers, 1939) [see Nelson (2016) for a discussion]. The automated feature extraction algorithm performed nearly all traditional analysis tasks, including identification of response onset and response peak and the calculation of numerical differences between response peak and response onset.

The automated feature extraction algorithm also selected relevant and comparison question pairs, such that for each repetition of the question sequences for the ZCT cases in the sample data the physiological responses to second and third relevant questions was paired with the physiological response to the preceding comparison question. For each re- cording sensor the automated algorithm paired the physiological responses to the first relevant question with the physiological response from either the preceding or subsequent comparison question by selecting the comparison question that produced the greater change in physiological activity.

Prior to feature extraction the automated algorithm scaled the recorded physiological data for visualization and performed some identification and rejection of data artifacts such as deep breaths in respiration activity, physical movement in the cardio data, and labile electrodermal responses that were unrelated to or untimely with the test question.

Dimensionless numerical measurements were obtained from the scaled physiological data from response onset to response end using a 15 second evaluation window (EW). Respiration excursion was measured as the mean of 1 second intervals from 0 to 14 seconds at a data rate of 30 sample per second, excluding 2 seconds prior to and sub- sequent to the verbal answer. Electrodermal and cardio responses were measured as the dimensionless difference of the maximum difference between the onset of a positive slope segment that began during the ROW and a subsequent peak of a positive slope segment. EDA and cardio response peak points were identified in the EW and were included in the extracted measurements if they occurred after the EW if the positive slope segment began during the ROW.

Measured segments for relevant and comparison questions were combined as the R/C ratio, which is the extracted relevant question measurement divided by the extract- ed comparison question measurement. R/C ratios were logged so that they produce values that are symmetrical around a mean

of zero.

Non-parametric integer scores were assigned to the physiological responses to analysis spots (i.e., relevant and comparison question pairs) using a three-position Likert (1932) type scale [+1, 0 -1]. Integer scores of positive sign value were assigned when the change in physiology was greater at the selected comparison question than the relevant question. Integer scores of negative sign value were assigned when the change in physiology was greater at the relevant question than the selected comparison question. Scores of zero sign (numerical zero) were assigned when the data were distorted from movement/activity artifact or were insufficient for scoring due to no response or due to a response onset prior to question onset, with no response onset during a defined response onset window (ROW) from question onset to five seconds after the verbal answer. Electrodermal and cardio responses were not used if they began during a .5 second latency period at question onset.

Non-parametric scores are intended to capture and express general information about the differences in response to relevant and comparison questions but do not pro- vide information that can be subject to linear assumptions about the degree of difference in response. However, some threshold constraints were used to prevent the assignment of numerical scores for segments of data that produced very little difference or extreme differences in responses to relevant and comparison questions. Threshold constraints were determined from optimization experiments with another a confirmed case sample used by Nel- son (2018a) and are shown in Table 2. Integer scores were assigned when the logged R/C ratio was within these constraints, and no score was assigned with the value was less than or greater than these constraint thresholds.

Table 2. Threshold constraints for non-parametric scores

-	Lower limit		Upper limit	
Sensor	Ratio	Logged ratio	Ratio	Logged ratio
Respiration	1.05/1.25*	.049/.223	1.5	.406
Electrodermal	1.05	.049	1000	6.908
Cardio	1.05	.049	1000	6.908

* An asymmetrical constraint 1.25 was used for the lower limit of respiration scores when assigning scores of + sign value. Optimization studies with other sampling data indicates that + scores are more likely to be negatively correlated with truth-telling without the asymmetrical constraint.

Data Reduction

Data reduction for each of the sample cases was accomplished by summing the numerical scores for respiration, electrodermal and cardio sensors for all presentations of each of the relevant test questions. In this way, a numerical subtotal score was obtained for each of the relevant questions. Subtotal scores were then summed to obtain a grand total score for each of the sample cases. Classifications of deception and truth-telling would be made with grand total and/or subtotal scores according to established polygraph decision rules.

Analysis

Sample cases were analyzed for correct decisions, errors and inconclusive results for the confirmed field cases and were calculated for test sensitivity, specificity, false-negative and false-positive errors, and inconclusive results for guilty and innocent sub-groups. In addition, the proportion of correct decisions was calculated for the guilty and innocent cases after excluding inconclusive results, along with the unweighted average of decision accuracy for the two groups. Positive predictive value and negative predictive values were also calculated as well as Detection efficiency coefficients. Accuracy indices were calculated for several conditions, including for three-position scores using traditional numerical cut- scores, after weighting the EDA scores, and using cut-scores selected from a multinomial reference distribution of weighted and unweighted three-position scores (Nelson, 2017, Nelson, 2018b).

Results

Results were tabulated for three-position scores using different decision rules and traditional numerical cut-scores. Results were also tabulated after doubling the value of all electrodermal scores. Finally, results were tabulated using cut-scores that were selected from multinomial reference distributions for both weighted three-position and unweighted EDA scores.

Decision Rules

Previous studies by Senter & Dollins (2003) have suggested that the choice of decision rules may play an important role in the effectiveness of polygraph classifications of deception and truth-telling. [See Nelson (2018c) for a discussion of different decision rules]. Table 3 shows the results three-position numerical scores for the n=100 confirmed field cases. Results are shown using the Federal ZCT Rule (FZR) with traditional numerical cut-scores (described earlier). Results for this study were also calculated using the Grand-Total Rule (GTR). Use of the GTR is a matter of summing all numerical scores and comparing the result to numerical cut-scores for deception or truth-telling (traditionally +6 and -6). Results were also calculated using the subtotal score rule (SSR), for which the lowest question subtotal (strongest indication of deception) is compared to numerical cut-scores for truth-telling (traditionally +3 or greater for all subtotals) or deception (traditionally -3 or lower for any subtotal). Also shown in Table 3 are of the sample cases using automated feature extraction and score assignment using the TSR with both traditional numerical cut-scores and the evidentiary cut-scores pro- posed by Krapohl (2005), as described earlier.

Table 3. Sample results for decision rules with three-position scores and traditional cut-scores (n=100).

	FZR	GTR	SSR	TSR*	EDR/TSR**
Sensitivity (deception)	.80	.56	.80	.80	.80
Specificity (truth-telling)	.32	.32	.06	.32	.48
False-negative Errors	<.01	<.01	<.01	<.01	<.01
False-positive Errors	.10	<.01	.10	.10	.10
Inconclusive-guilty	.20	.44	.20	.20	.20
Inconclusive-innocent	.58	.68	.84	.58	.42
Unweighted Accuracy	.88	>.99	.69	.88	.91
Unweighted Inconclusive rate	.39	.56	.52	.39	.31
Positive Predictive Value	.89	>.99	.89	.89	.89
Negative Predictive Value	>.99	>.99	.86	>.99	>.99
Detection Efficiency Coefficient	.79	.67	.79	.76	.82

* Results with the TSR are shown using the traditional numerical cut-scores.

** For comparison with Table 1, these evidentiary decision rules (EDR) results with the TSR are shown using asymmetrical cut-scores using by Krapohl (2005).

Test sensitivity in Table 3 is similar to that in Table 1, though specificity rates are lower for all decision rules. PPV and NPV for the three-position scores were greater than the results in Table 1, along with generally higher rates of inconclusive results. FN errors for the automated three-position scores was lower than in Table1. DEC was also increased for the FZR and TSR, and this can be attributed to the observed reduction in FN errors. It possible that these differences are due to the use of three-position vs seven-position numerical scores, though it is unknown what differences may be due to the use of automated feature extraction vs visual/subjective feature extraction.

PPV and NPV were highest for the GTR, though the inconclusive rate was also greatest for this decision rule. Light (1999) argued the rate of inconclusive cases for the GTR was un- acceptable for law enforcement use. However, the Light study is limited in scope – involving only confirmed guilty cases – and Table 3 shows that inconclusive rates are loaded on innocent cases. Also, Light did not include an evaluation of FP errors or the effectiveness of different numerical cut-scores. Inconclusive results for

guilty cases in Table 3 are greater than Table 1, and this is most likely attributed to differences in seven-position and three-position scores.

In the present study, using three-position scores and automated feature extraction, it is unclear whether the TSR provides any real advantage – in terms of classification accuracy – over the FZR. Results shown in Table 3 indicate a reduction of inconclusive results for innocent cases as a result of improved numerical cut-scores with the TSR. DEC was greatest for the TSR with the improved numerical cut- scores. This suggests the possibility that traditional numerical cut-scores for grand-total scores are inefficient for the reduced three-position scale.

Weighted Electrodermal Scores

Several publications have suggested that EDA data account for a larger portion of the diagnostic variance in test scores (Ansley & Krapohl, 2000; Harris, Horner & McQuarrie, 2000; Harris & Olson, 1994; Kircher, 1981, 1983; Kircher, Kristjianson, Gardner & Webb, 2005; Kircher & Raskin, 1988; Krapohl & McManus, 1999; Nelson, Krapohl & Handler, 2008; Raskin, Kircher, Honts & Horowitz, 1988) and contribute more information to effective conclusions about deception or truth-telling compared to the other recording sensors. To observe the differences in sample results with the three-position scores, EDA scores were doubled in value in the manner previously described by Krapohl and McManus (1999). Table 4 shows the sample results after weighting the electrodermal scores.

Table 4. Sample results for decision rules with n=100 confirmed field cases with traditional cut-scores and weighted electrodermal scores.

Specificity	FZR	GTR	SSR	TSR*	EDR/TSR**
Sensitivity (deception)	.92	.66	.92	.92	.92
(truth-telling)	.52	.54	.16	.54	.68
False-negative Errors	<.01	<.01	<.01	<.01	.02
False-positive Errors	.24	<.01	.24	.22	.14
Inconclusive-guilty	.08	.34	.08	.08	.06
Inconclusive-innocent	.24	.46	.60	.24	.18
Unweighted Accuracy	.84	>.99	.70	.86	.90
Unweighted Inconclusives	.16	.40	.34	.16	.12
Positive Predictive Value	.79	>.99	.79	.81	.87
Negative Predictive Value	>.99	>.99	>.99	>.99	.97
Detection Efficiency Coefficient	.87	.78	.86	.88	.90

* Results with the TSR are shown using the traditional numerical cut-scores.

** For comparison with Table 1, these evidentiary decision rules (EDR) results with the TSR are shown using asymmetrical cut-scores using by Krapohl (2005).

Weighting the EDA scores more than the other sensor score increased test sensitivity and specificity and reduced the occurrence of inconclusive results by 59% for the FZR, 29% for the GTR, and 35% for the SSR. The reduction of inconclusive results was 59% for the TSR with traditional cut-scores and 61% for the TSR using cut-scores that were suggested as optimal for evidentiary exams. The reduction of inconclusive results was greater for innocent cases than for guilty cases for all decision rules.

The GTR provided the greatest overall classification accuracy, though FN errors were low for all decision rules. The most obvious effect from weighting the EDA scores in Table 4, compared to Table 3, was a reduction of inconclusive results, along with increases in both test sensitivity to deception and specificity to truth-telling. The reduction of inconclusive results was greatest for the confirmed innocent cases. Interestingly, PPV was reduced for all models except the GTR which produced lower FN and FP error rates than other decision rules. DEC was improved for all decision rules and was

greatest for the TSR with improved cut-scores. This suggests that the selection of cut-scores may be an important consideration in the management and reduction of classification errors or inconclusive results.

Multinomial Cut-Scores

Multinomial Cut-Scores with Weighted Three-Position Scores

Weighted three-position scores were evaluated using cut-scores selected from multinomial reference distributions described by Nelson (2017). [See Nelson (2018b) for a discussion of how to use the multinomial reference distributions.] Multinomial cut-scores for weighted three-position of event-specific poly- graphs with three relevant questions, were as follows: grand total >= +3 for truthful classifications or <= -3 for deceptive classifications. Deceptive classifications were made using the subtotal scores when the grand total score was inconclusive if any subtotal score <= -7. The cut-score for subtotal scores was determined using a statistical correction for multiplicity effects to avoid the potential inflation of FP errors when using multiple subtotal scores for deceptive classifications. Results are shown in Table 5 for the combination of weighted EDA scores and multinomial cut-scores.

Table 5. Sample results (n=100) using weighted three-position scores with multinomial cut-scores.

	FZR	GTR	SSR	TSR
Sensitivity (deception)	.92	.88	.92	.92
Specificity (truth-telling)	.80	.80	.34	.80
False-negative Errors	.04	.04	.02	.04
False-positive Errors	.08	.06	.24	.08
Inconclusive-guilty	.04	.08	.06	.04
Inconclusive-innocent	.12	.14	.42	.12
Unweighted Accuracy	.93	.94	.78	.93
Unweighted Inconclusives	.08	.11	.24	.08
Positive Predictive Value	.92	.94	.79	.92
Negative Predictive Value	.95	.95	.94	.95
Detection Efficiency Coefficient	.94	.94	.85	.94

Use of multinomial cut-scores further improved test sensitivity to deception for the GTR, and improved test specificity to truth-telling for the FZR, GTR, SSR and TSR. Compared to the use of weighted electrodermal scores with traditional numerical cut- scores the multinomial cut-scores produced a reduction in inconclusive results by 50% for the FZR, 73% for the GTR, 29% for the SSR and 33% for the TSR. DECs for the multinomial cut-scores were consistently greater than for the traditional numerical cut-scores.

Not surprisingly, the SSR showed high sensitivity to deception, though not greater than any of the other decision rules, along with weaker specificity to truth-telling and a higher inconclusive rate that was loaded on innocent cases. Overall decision accuracy and DEC for the SSR was lower and inconclusive rates were higher than for other decision rules that included the use of the grand total score. This difference may be attributable to inherent multiplicity when using subtotal scores, and also to the smaller volume of information available to support decisions based on individual subtotal scores.

Multinomial Cut-Scores with Unweighted Three-Position Scores

Three-position scores were also evaluated using cut-scores obtained from a multinomial reference distribution for polygraph exams with three relevant questions and three to five charts. Multinomial cut-scores for the unweighted three-position scores were as follows: grand-total >= +2 for truthful classifications or <= -2 for deceptive classifications. Deceptive classifications were made using the

subtotal scores when the grand-total score was inconclusive if any subtotal score <= -6. The cut-score for subtotal scores was determined using a statistical correction for multiplicity effects to avoid the potential inflation of FP errors when using multiple subtotal scores for deceptive classifications. Results are shown in Table 6 for the combination of weighted EDA scores and multinomial cut-scores.

Table 6. Sample results (n=100) using unweighted three-position scores with multinomial cut-scores.

	FZR	GTR	SSR	TSR
Sensitivity (deception)	.88	.86	.80	.88
Specificity (truth-telling)	.78	.76	.30	.78
False-negative Errors	.04	.04	.02	.04
False-positive Errors	.04	.04	.10	.04
Inconclusive-guilty	.08	.10	.18	.08
Inconclusive-innocent	.20	.20	.60	.20
Unweighted Accuracy	.95	.95	.86	.95
Unweighted Inconclusives	.14	.15	.39	.14
Positive Predictive Value	.96	.96	.89	.96
Negative Predictive Value	.95	.95	.94	.95
Detection Efficiency Coefficient	.91	.90	.79	.91

Use of multinomial cut-scores improved the effectiveness of classifications with the unweighted three-position scores. Classification accuracy with the three-position scores using multinomial cut-scores was similar to that of the weighted multinomial model. However, both sensitivity and specificity were reduced slightly for the unweighted three-position scores. Compared to the weighted model the inconclusive rate for unweighted three-position scores increased by an average of 62% for all decision rules. The increase in inconclusive results was loaded on innocent cases. The unweighted three position scores also produced fewer false-positive errors. These differences can appear substantial when described as a percentage of change. The SSR underperformed relative to the other decision rules, with weaker test specificity and higher inconclusive results. Overall detection accuracy for the multinomial three-position model was high, though did not equal the effective- ness of the weighted three-position model.

Summary

This project involved the calculation of descriptive statistics for test accuracy, error and inconclusive rates as a function of different decision rules, structural weighting of sensor scores, and cut-scores. Results show that a number of procedural and field practice decision can have an important impact on the criterion accuracy of polygraph test results. Although PPV and NPV were consistently high for most conditions, differences in the rate of inconclusive results can be observed and this can directly affect the power of the test in terms of test sensitivity, specificity and error rates.

Results for the three-position scores, shown in Table 3, were similar for the FZR and TSR. However, there was a small reduction of inconclusive results, along with a corresponding increase in test specificity when using the TSR with cut-scores that were suggested as more optimal for evidentiary polygraph testing. Weighting the EDA scores, shown in Table 4, more than the other sensor scores increased test sensitivity and specificity, and reduced the occurrence of inconclusive results and average of 49%, compared to results from un- weighted three-position scores, for all decision rules. Use of multinomial cut-scores produced further reductions in the occurrence of inconclusive results along with further increases in DECs for all decision rules.

The combination of weighted EDA scores and multinomial cut-scores reduced the occurrence of inconclusive results by an average of 72% across all decision rules, compared to unweighted three-position scores and traditional numerical cut-scores. The reduction in inconclusive results was greatest for the GTR, which does not make use of subtotal scores, and for which the rate of inconclusive results with multinomial cut-scores was closer to that of other decision rules.

Field practitioners have provided anecdotal information suggesting that their observed rates of inconclusive rates are inconsistent with, and lower than, those in published studies. This is understandable because field practitioners, working at the level of individual cases, may be ethically justified in engaging in practices intended to resolve or reduce the occurrence of inconclusive results (e.g., con- ducting additional repetitions of the question sequence, or repeating an examination). In contrast researchers who work with samples of cases would be vulnerable to suggestions of manipulating a research outcome if they were to engage in such actions at level of some, though not all, individual cases. The result is that inconclusive rates in field practice may continue to be lower than those reported in published studies.

The most effective model in this analysis – illustrated by sensitivity, specificity, in- conclusive and DECs in Table 5 – was with the TSR using weighted EDA using multinomial cut-scores. Interestingly, accuracies for the FZR, and TSR were effectively identical for the weighted three-position scores with multinomial cut-scores, suggesting that the selection of cut-scores may be more important than the decision rules. Similarity of the DEC for the GTR, FZR and TSR provide further indication of this, and suggest that some previously re- ported conclusions about the GTR may have been unduly influenced by reliance on traditional numerical cut-scores that were inefficient for grand total scores.

This project, like all projects, is not without some limitations. The most obvious limitation is the small sample size (N=100). Though moderately sized for a project of this type, it is axiomatic that larger sample sizes are more easily viewed as comfortably approximating the population. However, sample size is not the only, or primary, consideration when attempting to understand the representativeness of a sample – for which random selection may be more important. This project, involving an archival sample, is necessarily precluded from any influence due to sampling methodology that is not presently expressed in the sampling data. It is also, necessarily de- pendent upon assumptions that the sampling data are in some way informative.

Another important limitation of this project is that no tests of statistical significance were completed. This was by design, as it was hoped that a descriptive approach to the statistical analysis might be of greater practical value to polygraph field examiners and program managers who may be more familiar and conversant with field practice policy decisions than multiple ANOVA. Future research should include a more complete analysis of the variance of the related effect sizes for polygraph decision rules, structural weighting coefficients for sensor scores, and cut-scores. Also, no statistical confidence intervals were included in this document, though informed readers can easily use a number of methods to calculate the confidence intervals of interest.

This project involved only a field sample of confirmed criminal investigation (event-specific) polygraphs and did not include a sample of multiple issue screening polygraphs. We suggest that some cautious generalization of these results is still in order. This is because of practical and important difference between event-specific diagnostic polygraphs and screening polygraphs that involve assumptions about the independence of multiple-is- sue screening questions. These assumptions are at best *convenience assumptions* because they assume independence in that different test items have no shared source of response variance – that whatever could influence responses to each item could have not affected any other item. As it happens, all polygraph questions within any examination will always have some shared source of response variance– in the form of the attention of the examinee. Both event-specific diagnostic and multiple-issue screening polygraphs will also be influenced by statistical multiplicity effects as determined by the selection of polygraph decision rules. For these reasons, a similar pattern of results can be expected for polygraph screening exams as is observed with this sample data.

Astute readers will note that this project does not attempt to discuss all possible methods of reducing inconclusive rates and improving polygraph test effectiveness. Some of those other methods may include: interviewing approaches, quality assurance activities, greater use of automation, use of the vasomotor sensor, recording additional charts, clarification of operational definitions, use of interview route-maps, refined target selection, and/or improvements in question formulation. All of these should remain as areas for continued research and development.

Results from this study point clearly to the fact that traditional numerical cut-scores are effective at producing a low FN rate but are burdened with unnecessarily weak test specificity to truth-telling and un-necessarily high rates of inconclusive results. An interesting observation that can be made about these results is that there was no advantage to the use of the SSR in terms of increased test sensitivity to deception, when compared to the other decision rules. The FN rate for the SSR was equal to that of the other decision rules for when using traditional cut-scores and was reduced from that of the TSR and FZR by 50% (.02 / .04), while the FP rate increased by a factor of 3 (.24 / .08) for the SSR. The inconclusive rate was larger for the SSR than for the FZR and TSR and was loaded on innocent cases. As shown in Table 3, the SSR was especially weak with unweighted three-position scores and traditional numerical cut-scores. The practical implication of these observations is that it may be difficult to justify the use of the SSR outside of polygraph screening con- texts, where it some over-prediction may be desirable or intended – and difficult to justfy the use of the SSR when using unweighted score and traditional cut-scores.

These results show clearly that optimization of field practices in each of these areas – decision rules, weighting of EDA scores and the selection of cut-scores – can provide important advantages to many, including polygraph field examiners, program managers, courts, legislators, researchers, and polygraph examinees. Further exploration is needed to better understand the utility functions in terms of economic values and operational costs associated with sensitivity, specificity, FN and FP errors, and inconclusive results rates. Greater reliance on statistical measurement theory can permit polygraph programs to refine their policies to better achieve their mission objectives and goals.

Inconclusive test results are likely to persist as a bane to polygraph field examiners, program managers, and others – including polygraph examinees. The availability of evidence-based procedural solutions that can reliably reduce the occurrence of inconclusive test results appears to be worthy of further attention and consideration.

References

Ansley, N. & Krapohl, D.J. (2000). The frequency of appearance of evaluative criteria in field polygraph charts. *Polygraph , 29,* 169-176.

Bradley, M. T. & Janisse, M.P. (1981). Accuracy demonstrations, threat, and the detection of deception: Cardiovascular, electrodermal, and pupillary measures. *Psychophysiology, 18,* 307-315.

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

Department of Defense (2006b). *Test data analysis: DoDPI numerical evaluation scoring system.* [Retrieved from http://www.antipolygraph.org on 3-31-2007].

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.

Harris, J. C. & Olsen, D.E. (1994). Polygraph Automated Scoring System. Patent Number: 5,327,899. U.S. Patent and Trademark Office.

Kircher, J. C. (1981). *Computerized chart evaluation in the detection of deception.* University of Utah. [Thesis],

Kircher, J. C. (1983). *Computerized decision making and patterns of activation in the detection of deception.* Doctoral dissertation, University of Utah, Salt Lake City. Dissertation Abstracts International, 44, 345.

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.

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

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

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

Kubis, J. F. (1962). *Studies in Lie Detection: Computer Feasibility Considerations. RADC-TR 62-205, Contract AF 30(602)-2270.* Air Force Systems Command, U.S. Air Force, Griffiss Air Force Base. New York: Rome Air Development Center.

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

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5-55.

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

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

Nelson, R. (2018a). Electrodermal signal processing: A correlation study of auto-centered EDA and manually-centered EDA with the criterion state of deception and truth-telling. *Polygraph & Forensic Credibility Assessment, 47 (1)*, 53-65.

Nelson, R. (2018b). Guide for how to use the ESS-multinomial reference tables in four steps. *APA Magazine, 51(2),* 78-89.

Nelson, R. (2018c). Practical polygraph: A survey and description of decision rules. *APA Magazine, 51(2),* 127-133.

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.

Raskin, D., 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.

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.

Summers, W. G. (1939). Science can get the confession. *Fordham Law Review, 8,* 334-354.

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), 70-7