# Survey of Computerized Polygraph Scoring Algorithms Using Kircher Features

**Raymond Nelson**

## Abstract

Published literature was surveyed for automated statistical classifiers that make use of a common set of physiological response features. These features were first introduced and described by researchers at the University of Utah during the 1980s, and have subsequently come to be described with the polygraph profession as *Kircher features*. These include the amplitude of increase for electrodermal and cardiovascular activity, along with a reduction of respiration activity. Constriction or reduction of vasomotor pulse amplitude can also be included. An interesting characteristic of these features, in addition to their statistical correlation with deception and truth-telling, they can be extracted from recorded time-series data both visually and via automated computer methods. Statistical classifiers based on these features include the following: Probability Analysis, a Rank Order Scoring System, an Objective Permutation method, a Bootstrap analysis method, the Empirical Scoring System/Multinomial, the Objective Scoring System (version 1 and 2), and the Objective Scoring System – version 3. Design characteristics of these analysis methods are summarized in the appendices.

## Introduction

Polygraph scoring algorithms, like all data analysis algorithms, consist of several fundamental operations or common functions. These include: feature extraction, numerical transformation and data reduction, use of some form of likelihood function, and structured rules or methods for interpretation and classification of output or results. Because all data analysis begins with feature extraction – the identification of useful and informative variation within the available data – development of knowledge about useful response features is an area of knowledge that be foundational to the development of varied methodological approaches when developing solution for the other, subsequent, functions within a data analysis method.

An example of this is a feature set that has come to be referred to as the *Kircher features*, (Kircher, 1981, 1983; Kircher & Raskin, 1988) as first described using this moniker by Krapohl and McManus (1999). In brief, these features consist of the primary signal for each of the traditional recording sensors: electrodermal phasic response amplitude, phasic increase in relative blood pressure, and reduction of respiration activity. Vasomotor activity can also be included in this feature set. Other researchers (Harris, Horner & McQuarrie, 2000; Kircher, Kristjansson, Gardner & Webb, 2004; Podlesny & Truslow, 1993; Rovner, 1986) have also shown the effectiveness of these response features.

Development of a data analysis method is a process of first specifying the desired output information –such as a statistical classification of deception or truth-telling – and then deconstructing the process of achieving that goal into a coherent set of assumptions and a reproducible series of functions. All algorithms consist of four essential functions, including: feature extraction, numerical transformation and data reduction, some form of likelihood function, and a set of rules or procedures used to interpret the test result. Of course, other process descriptions are also possible. However, these four basic functions can be generalized to nearly all data analysis methods, whether manual or automated, and whether based in traditional statistical classification and prediction methods or machine-learning/ artificial-intelligence.

Working backwards from through these functions, interpretation, in this usage refers to the process of translating a numerical and statistical result into conceptual information that may be useful or informative to persons not intimate with, or not involved in, the testing process or data analysis process. In other words, interpretation serves to answer the question: what does the test result actually mean? Categorical test results (i.e., positive or negative, and other allegorical terms) are the most simplistic form of interpretation and provide the smallest amount of detail about the scientific meaning of a test result – but often provide the most practical or actionable form of interpretation. In polygraph field practice, categorical results are the result of a procedural decision rule [1]. However, categorical results can be a source of misunderstanding and confusion when they are naively expected to be infallible. As a rule, scientific tests are not expected to be infallible. Although subject to dichotomous interpretation, all scientific test results are fundamentally probabilistic and therefore subject to inherent uncertainty. One of the main goals of any scientific test is to quantify and/or reduce the degree of uncertainty associated with a conclusion.

A *likelihood function* is a device used to obtain a reproducible statistical value for the observed test data. Likelihood functions can take many forms, including both empirical and mathematical distributions. A likelihood function can also be thought of as the parameters and formulae used to calculate a reference distribution. A simple and practical example of a likelihood function is a published table of values for a reference distribution. The most basic and simple form of likelihood function is a numerical cutscore at which a

[1]Polygraph decision rules make use of aggregated numerical information, such as grand total scores and subtotal scores. [See Nelson (2018) for a description of various decision rules including the grand total rule (GTR), two-stage rule (TSR), subtotal score rule (SSR), and others.]

categorical test result is selected. Numerical cutscores can be thought of as associated with some statistical likelihood that a test result is correct or incorrect.

Before any statistical value can be calculated, recorded data, if not already numerical, must be subject to *numerical transformation and data reduction*. Procedures for obtaining these numerical values can vary widely, and can include the use of physical measurement where applicable, likert scales, rank-ordering, use of ratios, z-scores or other mathematical or statistical values. Data reduction methods can also vary, and can include summation, averaging, weighted averaging, discriminate functions, log functions, resampling, and other methods. The functional objective of data transformation and numerical reduction is to transform recorded test data to a set of useful numerical values, and to reduce those values to a small set of numbers for which a statistical value can be obtained.

Before any numerical or statistical results can be calculated, useful response features must be extracted from the recorded data. *Feature extraction* is the beginning of any algorithmic or procedural method for data analysis. All data are a combination of signal and noise. Feature extraction is the process of identifying and isolating the response information of interest, including the identification of response onset and response end. Under ideal circumstances the ratio of signal to noise is very high, and it is very easy to isolate useful signal information from useless noise. Feature extraction research is foundational to the development of solutions for all subsequent analytic functions[2].

## Method

Published literature was surveyed for descriptions of statistical classifiers based on physiological response features described by researchers at the University of Utah (Kircher & Raskin, 1988). Design and development characteristics were enumerated for these methods. Information was sought for the type of decision method and type of statistical classifier, along with methods for numerical transformation and data reduction – including the selection of selecting relevant-comparison question pairs where applicable. General methods for the development of a statistical likelihood function are described for each analysis algorithm, Finally, procedural rules for interpretation or classification of deception and truth-telling are described for each analysis method, as applicable to single issue and multiple issue polygraph examinations.

## Results

Seven different statistical classifiers were found in the published literature. These include: Probability Analysis (PA; Kircher & Raskin, 1988), a Rank Order Scoring System (ROSS; Honts & Driscoll, 1987; 1988), a Bootstrap Analysis Method (BAM; Honts & Devitt, 1992), the Objective Scoring System – versions 1-2 (OSS 1-2: Krapohl, 2002; Krapohl & McManus, 1999), an Objective Permutation Scoring method (OPS; MacLaren & Krapohl 2003), the Objective Scoring System version 3 (OSS-3; Nelson, Krapohl & Handler, 2008), and the Empirical Scoring System – Multinomial (ESS/ESS-M; Nelson, 2017a; Nelson, Krapohl & Handler, 2008).

Information of interest to this survey, in addition to the use of Kircher features and associated recording sensor, included the following: decision model, statistical classifier, numerical transformation, method or procedure for selecting RQ and CQ analysis pairs, data reduction method, type of likelihood function, and procedural decision rules. Type of decision model refers to the overall method by which a classification is achieved; this can include the use of a z-test, gaussian-gaussian signal discrimination, simple Bayes, or other method. A number of types of statistical classifiers were observed, including, p-values, posterior odds, point and cutscore comparisons that are mapped to TN and TP rates, and other methods. Polygraph field examiners who desire to

[2]For simplicity, this discussion ignores the series of important functions prior to data analysis, including sensor development, stimulus development, test administration, signal processing, and data recording.

better understand differences between manual and automated scoring methods may be interested in the variety of methods employed for the selection and comparison of RQ and CQ value pairs. Methods for numerical transformation included rank transformations, difference scores, z-scores, and other methods. Strategies for reducing sensor and question subtotals to grand total scores included summation, averaging and weighted averaging. A variety of types of likelihood functions were observed, including empirical distributions, multinomial distributions, bootstrap distributions, permutations, and other methods. Decision rules included the use subtotal scores for multiple issue exams, grand total scores for single issue exams and other procedural solutions. Design and development characteristics for each of these analysis methods is shown in Appendices A-G.

## Conclusion

Kircher features have been a useful and effective solution to the challenges of polygraphic feature extraction since the 1980s. The existence of an easily identified and easily used feature set has facilitated the study and development of a variety of types of statistical classifiers. These features were first described in the development and of the Probability Analysis algorithm (Kircher & Raskin, 1988). At this time, seven different computer algorithms can be found in the published literature that make use of this common polygraph feature set. Many of these methods are available in commercial and professional products in use by polygraph field practitioners.

Because all of the surveyed analysis methods are based on Kircher features, they all included similar recording sensors. Some differences in signal processing may exist for different polygraph instruments. However, signal processing differences are beyond the scope of this project. A number of different types of statistical classifiers are included in the surveyed analysis methods. These include maximum likelihood estimation, linear discriminate analysis, gaussian-gaussian signal detection/discrimination, permutation and bootstrapping methods, and rank transformations. Some methods make use of simple Bayesian classifiers, for which the posterior result may be thought of pragmatically as a probability of deception or truth-telling.

Vasomotor response data, although it can be thought of as one of the Kircher features, is not included in most these analysis methods. Although information on vasomotor activity can be found in the published literature – including publications by some developers of available scoring algorithms – it has not been included in the structural model of the published and available scoring methods. Reasons for this have not been completely discussed. However, it can be assumed that vasomotor data would more likely have been included if it had improved the effect sized of the published structural models. It can therefore be hypothesized that vasomotor data, though perhaps correlated with the criterion of interest, may not have improved the structural models described in in the published literature. Addition of vasomotor data would require a sufficient basis of data with which to re-develop the various likelihood functions and study the resulting effect sizes. One analysis method, the ESS-M (Nelson, 2017a) does include a likelihood function that can include vasomotor data. However, published information does not show any difference in effect size when including the vasomotor information (Nelson, 2017b). A more complete understanding of the potential vasomotor response data and effect sizes for automated scoring methods will require replication and extension of these algorithm methods, in addition to the recalculation or redevelopment of associated likelihood functions.

Limitations of this project are several, and include the fact that this project is intended only to provide a descriptive summary of design characteristics of these different methods. No mathematical or procedural description of the identified scoring methods is included in this report. Another limitation of this survey is that it does not include other computerized analytic methods that make use of other scoring features. Other analysis methods may exist in publication, including methods that rely on proprietary and boutique feature extraction methods, and response features that are less familiar or intuitive for field polygraph practitioners – for example, spectral response features. This project is limited to methods

that exist in open publication, and does not include algorithms that are subject to proprietary or intellectual property restriction. Other research should address the need for information on those methods. Finally, no information on effect sizes is described for the algorithms included in this report. Other research should address the need for information on effect sizes – including sensitivity, specificity, false-positive, and false-negative rates, and associated errors of measurement. Future research should further investigate potential advantages to the various solutions to the series of challenges inherent to automated statical data analysis and classification.

This project is a brief description of conceptual, albeit non-technical, information that may be useful to readers who want an introduction to the topic of polygraph computer scoring algorithms, along with an introduction to the breadth of activity in this area throughout the past 35 or more years. In addition to the fact that the availability of a useful set of known response features has enabled a variety of researchers to study the application different statistical methods to classification of deception and truth-telling, Kircher features have the advantage that they can provide some intuitive understanding to field polygraph examiners who desire to understand what details of the recorded physiology is included in the analysis. Indeed Kircher feature can be extracted either manually or via automation. It is hoped that this information is useful or informative to those interested in polygraph data analysis algorithm development, and to field polygraph professionals who wish to more fully understand differences between various analysis methods and traditional manual scoring procedures.

## References

Harris, J. Horner A. & McQuarrie, A. D. (2000). *An Evaluation of the Criteria Taught by the Department of Defense Polygraph Institute for Interpreting Polygraph Examinations*. Department of Defense Polygraph Institute report No DoDPI00-R-0007

Honts, C. R. & Devitt, M. K. (1992). *Bootstrap decision making for polygraph examinations*. Report number DoDPI92-R-0002. Department of Defense Polygraph Institute report No DoDPI92-R-0002. Reprinted in Polygraph, 3, (1), 1-47.

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*, 1-15.

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

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., Kristjansson , S. D., Gardner, M. D. & Webb, A. (2004). *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. (2002). Short report: Update for the objective scoring system. *Polygraph, 31*, 298-302.

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

MacLaren, V. V. & Krapohl, D. J. (2003). Objective Assessment of Comparison Question Polygraphy. *Polygraph, 32*, 107-126.

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.

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

Nelson, R. (2017b). Updated numerical distributions for the Empirical Scoring System. An accuracy demonstration with archival datasets with and without the Vasomotor Sensor. P*olygraph & Forensic Credibility Assessment, 46* (2), 116-131.

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

Nelson, R. (2019). Literature survey of structural weighting of polygraph signals: why double the EDA? *Polygraph & Forensic Credibility Assessment, 48* (2), 105-112.

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.

Rovner L. I. (1986). The accuracy of physiological detection of deception for subjects with prior knowledge. *Polygraph, 15*(1), 1–39

## Appendix A. Probability Analysis

| Probability Analysis (Kircher & Raskin, 1988) | |
|---|---|
| Sensors | Respiration (thoracic and abdominal), Electrodermal, Cardiovascular |
| Response Features | Reduction of respiration activity, electrodermal amplitude of phasic response, cardiovascular phasic increase in relative blood pressure. |
| Decision model | The algorithm computes a discriminate function that serves as a statistical classifier – a statistical value value for which decision cutpoints can make categorical classifications of deception or truth-telling. |
| Statistical classifier | Bayesian analysis using a likelihood function obtained via discriminant analysis. Results can be thought of as a posterior probability of deception or truth-telling. |
| CQ Selection | For each sensor, between chart mean for all RQs is compared to the between chart mean for all CQs. |
| Numerical transformation | Numerical values transformed to z-scores for each sensor using combined RQs and CQs. |
| Data reduction | Z-scores are averaged between-charts for the individual sensors for all RQs, and for all CQs. Sensor z-scores are then combined using a structural weighting function that was obtained using linear discriminate analysis. |
| Likelihood function | Two likelihood formulas are used to calculate complimentary likelihood values for deception and truth-telling. |
| Decision rules – single issue | GTR |
| Decision rules – multiple issue | none |
| Comments | Structural coefficients are available from the developers and also from replication studies. Publication describe the application of PA to single issue exam formats. Application of the PA algorithm to multiple issue exams may require a change from the aggregation of RQs and CQs both within and between charts. Advantages of separate within-chart and between-chart transformation schemes have not been fully described in publication, however subsequent algorithm have shown the application of these to multiple issue exams. For this reason it may be possible to adapt PA to multiple issue test formats. |

## Appendix B. Rank Order Scoring System

| ROSS (Honts & Driscoll, 1987, 1988) | |
|---|---|
| Sensors | Respiration (thoracic and abdominal), Electrodermal, Cardiovascular, Vasomotor |
| Decision model | ROSS decision model is similar to a gaussian-gaussian signal discrimination method, using empirically derived summed rank distributions for guilty and innocent cases. |
| Statistical classifier | Statistical classifiers are empirically derived TP, TN, FP, and FN rates. |
| Response Features | Reduction of respiration activity, electrodermal amplitude of phasic response, cardiovascular phasic increase in relative blood pressure. |
| CQ Selection | CQs and RQs are not paired for analysis as in tradition polygraph scoring. Instead, rank order analysis begins with the assignment of integer rank scores to all test stimuli, including all CQs and all RQs together, within each recorded chart |
| Numerical transformation | Integer scores are assigned to rank order variance of RQs and CQs for each sensor, within each chart. Rank scores are assigned in reverse order, wherein the response with the greatest change in physiology is assigned a rank value equal to the total number of RQs and CQs, and the smallest response receives a rank value of 1. |
| Data reduction | Rank values are summed for all RQs for all charts, and also for all CQs for all charts. A CQtotal – RQtotal = RankDifference score is then calculated. |
| Likelihood function | Empirical distributions can be calculated for RankDifference scores, and numerical cutscores can be selected to achieve desired effect sizes. |
| Decision rules – single issue | GTR |
| Decision rules – multiple issue | SSR |
| Comments | Rank order transformations are a common non-parametric solution, and can sometimes optimize robustness with messy and difficult data with some potential cost due to the granularity of rank values. Each rank value is obtained by comparing each response to all other response (all other RQs and CQs), which may complicate assumptions of independent RQ variance for multiple issue exams. |

## Appendix C. Bootstrap Scoring System

| Bootstrap Scoring System (Honts & Devitt, 1992) | |
|---|---|
| Sensors | Respiration (thoracic and abdominal), Electrodermal, Cardiovascular, Vasomotor |
| Decision model | Z-test of an observed CQ-RQ difference using a bootstrap null distribution. |
| Statistical classifier | P-value, indicating the likelihood of obtaining a score equal to or more extreme than the observed score under the null hypothesis that there is no difference between CQ and RQ value. |
| Response Features | Reduction of respiration activity, electrodermal amplitude of phasic response, cardiovascular phasic increase in relative blood pressure. |
| CQ Selection | Each RQ is paired with the preceding CQs to calculate CQ – RQ difference values after transforming all RQ and CQ values to z-Scores. |
| Numerical transformation | For each sensor, presentations of all RQs and CQs, for all combined charts, are transformed to z-Scores. In this way all scores for all sensors and all charts have a common scale value that can easily subject to bootstrapping. |
| Data reduction | CQ – RQ = Z-difference scores were calculated for the z-Scores. Z-Difference scores are hypothesized to be loaded at greater than zero values innocent subjects and less and zero for guilty subjects. Z-difference scores are aggregated via summation to obtain a single Z-difference score for an exam. |
| Likelihood function | A null distribution is calculated for each exam by combining all RQ and CQ z-Scores, for all charts and all sensors, into a single vector, and then bootstrapping a null distribution (random sampling with replacement) while arbitrarily assigning values as CQ or RQ. |
| Decision rules – single issue | GTR |
| Decision rules – multiple issue | NA |
| Comments | This method was described using single issue exams with an equal number of RQs and CQs, but could be adapted to multiple issue test formats and test formats with unequal numbers of RQs and CQs. |

## Appendix D. Objective Scoring System (OSS/OSS-2)

| OSS/OSS-2 (Krapohl & McManus, 1999; Krapohl, 2002) | |
|---|---|
| Sensors | Respiration (thoracic and abdominal), Electrodermal, Cardiovascular, |
| Decision model | Gaussian-Gaussian signal-discrimination model (signal detection classifier applied to a signal discrimination task). |
| Statistical classifier | P-value, indicative of the likelihood of the observed test statistic under the distribution represented by the training data confirmed as opposite of the selected (deceptive or truthful) classification. |
| Response Features | Reduction of respiration activity, electrodermal amplitude of phasic response, cardiovascular phasic increase in relative blood pressure. |
| CQ Selection | Each RQ is compared to the preceding CQ |
| Numerical transformation | R/C ratios are transformed to integer scores using a distribution of uniform septile bins. OSS-2 7-position scores [-3, -2, -2, 0, +1, +2, +3] differ from tradition polygraph 7 position scores in that the range of OSS-2 scores can occur with equal likelihood, whereas traditional 7-position scores are loaded near 0 with scores further from 0 occurring less frequently. |
| Data reduction | Integer scores are aggregated via summation for each RQ, for all sensors and all charts. RQ subtotals are then summed for a grand total score. Because integer scores are aggregated via summation it makes no difference whether sensor scores are summed first between charts or within-charts. |
| Likelihood function | OSS-2 reference tables are empirically derived |
| Decision rules – single issue | GTR |
| Decision rules – multiple issue | none |
| Comments | OSS-2 likelihood functions (reference tables) available for single issue polygraph exams with 3 RQs and 3 charts. The summative design means that the likelihood function may be less robust with missing and artifacted data, and may be overloaded when more than three charts are used, and may become biased with test formats with unequal numbers of RQs and CQs. OSS likelihood functions have not been published for multiple issue exams, OSS-1 and OSS-2 began as manual scoring protocols, for which the structure and procedures were sufficiently structured and unambiguous that they led easily to automation. A result of this is that OSS and OSS-2 are now defacto automated analysis methods. |



## Appendix E. Permutation Scoring System

| Permutation Scoring System (MacLaren & Krapohl, 2003) | |
|---|---|
| Sensors | Respiration (thoracic and abdominal), Electrodermal, Cardiovascular, Vasomotor |
| Decision model | A simple bayesian classifier using odds form of Bayes' theorem, where p/(1-p) values obtained are from a permutation of uniform 7 position integer scores. |
| Statistical classifier | Use of Bayes' theorem means that results can be thought of in practical terms as a posterior probability of deception or truth-telling. |
| Response Features | Reduction of respiration activity, electrodermal amplitude of phasic response, cardiovascular phasic increase in relative blood pressure. |
| CQ Selection | RQs are paired with preceding CQs. |
| Numerical transformation | A ratio is calculated for each RQ/CQ pair, after which two sets of integer scores are assigned using two distributions of uniform septile bins that were calculated from confirmed guilty and innocent cases. |
| Data reduction | Two sets of integer scores are summed to obtain two grand total scores (guiltyTotal and innocentTotal) which are then compared to a PSS likelihood function. |
| Likelihood function | PSS likelihood function is the permutation of all possible 7 position scores if they are not systematically associated with guilt or innocence. The exact distribution includes 6.57 x 10^22 possible combinations. It can be calculated using a combinatoric formula, and can be easily approximated via simulation. |
| Decision rules – single issue | GTR |
| Decision rules – multiple issue | none |
| Comments | PSS was developed with examination consisting of three presentations of a question sequence that includes three RQs and 3 CQs. Adapting the PSS method to multiple issue exams, and to test formats with two or four RQs, and with four or five presentations requires available confirmed case data to calculate the uniform septile distributions, in addition to recalculation of the permutation likelihood function. |

## Appendix F. Objective Scoring System – version 3

| OSS-3 (Nelson, Krapohl & Handler, 2008) | |
|---|---|
| Sensors | Respiration (thoracic and abdominal), Electrodermal, Cardiovascular |
| Decision model | Gaussian-Gaussian signal-discrimination model (signal detection classifier applied to a signal discrimination task). |
| Statistical classifier | P-value, indicative of the likelihood of the observed test statistic under the distribution represented by the training data confirmed as opposite of the selected (deceptive or truthful) classification. |
| Response Features | Reduction of respiration activity, electrodermal amplitude of phasic response, cardiovascular phasic increase in relative blood pressure. |
| CQ Selection | Mean CQ is compared to each RQ. |
| Numerical transformation | Log R/C ratios standardized to the training data. Sensor scores are standard scores (mean = 0, standard deviation = 1), from -3 to +3. Standardized log R/C ratios indicated the number of standard deviations an observed response is above or below the mean of the training data when guilty and innocent cases are combined. These scores are intuitively similar to the notion of 7 position polygraph scores, but with decimals. |
| Data reduction | Grand mean is the mean of between-chart RQ scores. Between-chart RQ scores are the means of within-chart weighted mean sensor scores. Sensor scores are standardized log R/C ratios. Sensor weighting coefficients obtained through linear discriminate analysis, and can also be calculated via logistic regression, bootstrapping and other methods with little difference in the resulting weighting function. |
| Likelihood function | OSS-3 reference tables are empirically derived for confirmed guilty and innocent cases. |
| Decision rule – single issue | GTR, TSR |
| Decision rule – multiple issue | SSR, OSS-3 Screening rule uses the K-W ANOVA method to evaluate differences and similarities between RQs to reduce the occurrence of inconclusive with multiple-issue exams (based on an assumption that truth-telling to all RQs will result in no significant differences in RQ scores). |
| Comments | OSS-3 was intended extend available knowledge from OSS-1 and OSS-2 to a wide variety of examination formats, including single and multiple issue exams with 2, 3, and 4 RQs, and 3, 4, or 5 charts. OSS-3 was designed to be robust with some missing and artifacted data, and to use the 2nd of any repeated questions within a chart. Data reduction via averaging means that OSS-3 likelihood functions are more readily applicable to exams with 3, 4 or 5 presentations of the question sequence, and the log(RQ/CQmean) transform is applicable to test formats with unequal numbers of RQs and CQs. The algorithm also includes capabilities to mark artifacted and unusable segments for exclusion from analysis. Artifacted segments can be analyzed, using a test of proportions, to make inferences about their cause, whether systematic or random. |

## Appendix G. Empirical Scoring System (ESS/ESS-M)

| ESS/ESS-M (Nelson, 2017a; Nelson, Handler & Krapohl, 2008) | |
|---|---|
| Sensors | Respiration (thoracic and abdominal), Electrodermal, Cardiovascular, Vasomotor |
| Decision model | ESS relies on a gaussian-gaussian signal discrimination model. ESS- is a simple bayesian classifier. ESS can also be studied and used with traditional/Federal cutscores to achieve empirically studied effect sizes. |
| Statistical classifier | ESS uses a p-value, obtained from an empirical distribution, that describes the likelihood of the observed data under a specified hypothesis. Cutscores for ESS were selected empirically, to constrain FN and FP errors to desired alpha levels. ESS-M results can be expressed as the posterior odds of deception or truth-telling. Use of Bayes' theorem means that results can be thought of in practical terms as a posterior probability of deception or truth-telling. ESS-M also provides the lower-limit odds of the 1-alpha posterior credible interval for deception or truth-telling – indicative of the likelihood of obtaining a similar categorical result upon repetition of the test procedure. |
| Response Features | Reduction of respiration activity, electrodermal amplitude of phasic response, cardiovascular phasic increase in relative blood pressure. Can optionally include vasomotor reduction of pulse amplitude. |
| CQ Selection | RQs are paired with CQs according to traditional procedures used by field examiners for each different polygraph test format. In general, RQs are compared to the preceding or subsequent CQ with the greater change in physiological activity whenever possible, and with the preceding CQ when two CQs are not available. |
| Numerical transformation | Integer scores are assigned by comparing differences in response magnitude for RQ and CQ pairs. Question pairs can be used naively or subject to optimization coefficients to reduce scores that may occur due to spurious or random noise. EDA integer scores are doubled prior to summation, so that the structural contribution of EDA data is greater than for other sensors. [See Nelson, 2019.] |
| Data reduction | Integer scores are summed for each RQ, for all sensors, between charts. RQ subtotals are then summed to obtain a grand total score. Summation via this process means that values are available for single issue and multiple issue test formats. |
| Likelihood function | ESS likelihood functions were calculated empirically using only respiration, EDA and cardiovascular sensors. ESS-Multinomial likelihood functions are calculated mathematically using the analytic theory of the CQT, and available for single issue and multiple issue polygraph exams with 2, 3, and 4 RQs with 3, 4 or 5 charts, including respiration, EDA, and cardiovascular sensors, in addition to the optional vasomotor sensor. Likelihoods for traditional cutscores are the empirical TP, TN, FP and FN rates. |
| Decision rules – single issue | GTR, TSR, FZR |
| Decision rules – multiple issue | SSR |
| Comments | ESS and ESS-M were introduced as manual scoring methods. Research publications have made use of fully automated ESS and ESS-M models. |