Accuracy Effects for ESS and Three-Position Scores of Federal ZCT Exams

Using the Grand Total Rule with Traditional/Federal and Multinomial Cutscores

Raymond Nelson

Abstract

This project is a comparison of accuracy effects for the ESS and three-position scores using traditional numerical cutscores and multinomial cutscores. Effects studied include test sensitivity, specificity, false-positive and false negative error rates, in addition to positive-predictive-value, negative-predictive value, the proportions of correct classification for guilty and innocent cases and the unweighted mean of correct and inconclusive cases. An archival samples of n=100 confirmed field cases using the Federal ZCT format were used, permitting intuitive comparison with previously published effects. A second sample of n=60 confirmed field cases using the Federal ZCT format was also included in the analysis. Responses were extracted from the recorded data and scores were assigned via an automated ESS algorithm that was designed to closely approximate the feature extraction process used by human experts when manual scoring polygraph data. ESS scores were then converted to three-position scores. A parametric bootstrap was used to calculate statistical confidence intervals at the .025 and .975 percentiles, and to estimate the variance of observed effects. A mixed-effects ANOVA procedure was used to evaluate the four treatments of the sample cases: ESS and three-position scores with traditional and multinomial cutscores. Accuracy for the four treatments when excluding inconclusive cases was similar for positive-predictive-value, negative-predictive-value, and the proportions of correct classifications excluding inconclusive results. Use of multinomial cutscores contributed to a statistically significant reduction of inconclusive results, and statistically significant increases in test sensitivity to deception and test specificity to truth-telling for both ESS and three-position scores. None of classifications of the individual cases were observed to change from deceptive to truthful or from truthful to deceptive for any of the four treatments with either of the two archival samples.

Introduction

The Empirical Scoring System (ESS; Nelson et al., 2011) is an evidence-based, standardized and statistically referenced method for the analysis of psychophysiological detection of deception (PDD) test data. The ESS can be thought of as a modification of the Federal three-position scoring method (National Center for Credibility Assessment, 2017), which can itself be thought of as a modification of the Federal seven-position scoring method. Previous studies on the ESS indicate that it provides accuracy effects similar to the Federal seven-position method with the practical reliability and intuition of the three-position scoring method. Results for the ESS were first described in a comparison of accuracy effects (Nelson, Krapohl & Handler, 2008) for polygraph examiner trainees with those from experienced examiners. The ESS was updated (Nelson, 2017a; 2017b) to make use of a Bayesian classifier and cutscores that were obtained from a multinomial reference distribution that was calculated under the analytic theory of PDD testing (Nelson, 2106). Multinomial cutscores and reference distributions were subsequently calculated for the three-position scoring method (Nelson, 2018a). ESS is widely used by polygraph examiners across the U.S. and worldwide, including professionals in private practice and in municipal, state and federal law enforcement/investigation agencies.

The ESS differs from the three-position scoring method in three main ways. First, ESS scores are assigned while doubling the value of all EDA scores. This is so that electrodermal scores are weighted in a manner that approximates the structural and statistical functions described in the scientific literature on PDD test data analysis and computer algorithm development Nelson (2019). A second important difference is that the ESS makes use of statistically referenced numerical cutscores in lieu of traditional cutscores that were derived heuristically for the seven-position scoring method. Another third difference is that different agencies have implemented the ESS with different decision rules, according to operational and mission objectives.

Decision rules define the structured procedures used to interpret and parse the categorical test result from numerical and statistical information. [Refer to Nelson (2018b) for a literature summary and description of polygraph decision rules.] Commonly used decision rules in PDD field practice include the grand-total-rule (GTR), two-stage-rule (TSR), and the Federal Zone Rule (FZR) and the subtotal-score-rule (SSR). Among these the GTR has been shown consistently to provide the highest level of classification accuracy for single issue exams, while the SSR has been regarded by many polygraph agencies and field examiners as the optimal decision rule for multiple issue screening exams.

This project is a comparison of accuracy effects for ESS using the GTR with traditional cutscores and multinomial cutscores. Two archival samples were used in this project, permitting intuitive comparison with previously published accuracy effects. Also studied were classification accuracy effects with three-position scores.

Methods and Materials

Data

Data for this project were a confirmed field sample of n=100 exams that were conducted using the Federal Zone Comparison Test (FZCT) format (Department of Defense, 2006). This sample was previously used by Krapohl and Cushman (2006) with Federal seven-position scores, and later by Nelson, Krapohl and Handler (2008) in an early study on the ESS. Sample cases were conducted by a variety of federal, state, and municipal law enforcement agency and were subsequently included in the confirmed case archive at the Department of Defense Polygraph Institute (now the National Center for Credibility Assessment). The FZCT is a three-question, event-specific test format, that is recognized as among the most useful test formats for the investigation of criminal incidents. All cases consisted of three iterations of a question sequence that included three relevant-questions (RQs) and three comparison-questions (CQs) in addition to other procedural questions that are not subject to numerical or statistical analysis. Field polygraph examiners refer to the repetitions or iterations of the question sequence as “charts,” with reference to old-time polygraph instruments that plotted physiological data through capillary ink pens onto rolled chart paper. Human expert, when scoring the sample cases manually, have described some of the sample cases as challenging. Although perhaps not ideal, use of this same sample data can provide practical and intuitive understanding of differences in accuracy effects for different test data analysis methods. [Refer to Nelson (2015) and Department of Defense (2006) for general information on the comparison question test and how the sample cases were conducted.]

All cases included the standardized array of PDD sensors, for which physiological responses and numerical scores would be extracted, including: upper and lower respiration sensors, an electrodermal activity sensor, and cardiovascular activity sensor. Acquired knowledge pertaining to the FZCT format, in addition to basic principles and procedures, have been generalized to other PDD formats including single-issue and multiple-issue use-cases with two, three and four RQs. This sample was used in the initial study and development of empirical reference distributions for the ESS, and was subsequently used in an accuracy demonstration of the multinomial update to the ESS-M (Nelson, 2017b).

A second archival sample was obtained, consisting of n=60 confirmed field exams using the FZCT format. These exams were also included in the DoDPI (now NCCA) confirmed case archive. This sample was previously used as the holdout validation sample in the development of the OSS scoring algorithms (Krapohl & McManus, 1999, Krapohl, 2002, Nelson Krapohl & Handler, 2008), and was also used in a study of manual scoring with the Federal seven-position and ESS scoring methods. All exams in the second dataset consisted of three iterations of a question sequence that included three RQs and three CQs in addition to other procedural questions. Similar to the first archival sample, these examinations were conducted by a variety of municipal, state and federal law enforcement agencies.

Analysis

Sample data were analyzed using an automated version of the ESS. All tests data analysis methods – regardless of whether polygraph or other form of test – will consist of similar functions, feature extraction, numerical transformation and data reduction, use of some form of likelihood function or statistical classifier, and structured procedures for the interpretation and classification of result. Feature extraction refers to the identification of useful or diagnostic information in the recorded test data, and the extraction or separation of this information from other non-useful information or noise. Numerical transformation, when manually scoring polygraph test data, is the conversion of observed physiological responses to numerical values – using a system of [+, 0, -] integers. The simplest form of likelihood function is a numerical cutscore for which classification effects can be known, including true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) outcomes). Another form of likelihood function will map or obtain numerical cutscores to either an empirical or a theoretical reference distribution – both of which are available in publications for the ESS for which a multinomial reference distribution can be calculated under the analytic theory of PDD testing. Regardless of the form of likelihood function, parsing a categorical result from numerical and statistical test data requires the use of a structured decision rule. The automated ESS was designed to replicate objectively the procedures used by human experts when manually scoring PDD test data, including feature extraction, selection of RQ and CQ analysis spots, assignment and aggregation of integer scores, numerical cutscores, and decision rules.

Signal processing

Time-series data for all sample cases were exported to the NCCA ASCII format (Editorial Staff, 2019) and imported into the R Statistical Computing Language and Environment (R Core Team, 2019) to complete the signal processing, feature extraction, and data analysis. Signal processing of the digitized data was completed at a data rate of 30 cycles per second (cps) for all recorded sensors. Respiration data were subject to a smoothing filter, consistent with previous publications, consisting of a first-order Butterworth type low-pass filter (Butterworth, 1930) with a corner frequency of .886Hz (equivalent to a moving average filter with a .5 second window). Smoothing filters of this type have been shown to improve the correlation and diagnostic coefficients obtained from respiration data (Nelson & Handler, 2012).

All examinations were conducted using Axciton computerized polygraph systems that included a hardware-based high-pass filter (auto-centering EDA) option in addition to the manually-centered EDA option. Discussion with field practitioners revealed a common belief that field practices favored the use of manually centered EDA at the time the examinations were conducted. This may have been a result of the fact that engineering specifications of hardware-based high-pass filter of old-time analog polygraph instruments was largely undocumented as to the corner frequencies or time-constants of the filter design. Similarly, the corner frequency and time-constant for the Axciton computerized polygraph system has been described in previous publications as unknown (National Research Council, 2003). No information was captured or recorded regarding the selection of the EDA mode for the sample cases. A consequence of this is that it is possible that some of the sample cases were recorded using the hardware-based high-pass filter, and no attempt was made to determine the EDA mode through visual inspection. For this reason, no high-pass filter was used for the EDA data, and signal processing for EDA data was limited to the reduction of high-frequency noise through a first-order Butterworth type low-pass filter with a corner frequency of .886Hz.

Cardio data includes both low-frequency blood-pressure information and higher frequency pulse rate information. Because of the need to avoid altering or disrupting diastolic and systolic cardio peaks, cardio data was not subject to additional signal processing or smoothing.

The NCCA ASCII specification makes use of dimensionless units that are not associated with a standardized physical measurement. For this reason, scaling of the data has no effect on analytic results for individual cases or for this analysis.

Feature extraction

Feature extraction was accomplished using an automated procedure intended to replicate that used by human experts when manually scoring PDD data. Physiological reactions were evaluated using a 15 second evaluation window (EW) for all recording sensors. This EW is thought to be sufficient to observe most physiological responses to test stimuli and is regarded as somewhat robust for persons with common difficulties with sustained attention. For EDA and cardio data, a response-onset-window (ROW) was defined as from stimulus onset to five seconds after the point of verbal answer, or five seconds after stimulus offset if there was no recorded verbal answer.

Respiration feature extraction. For respiration data, information was excluded from the feature extraction for 1.5 seconds prior to and 1.5 seconds following the recorded point of verbal answer. This was to avoid the inclusion of commonly occurring answering distortions in the respiration feature extraction. Respiration data were measured using the respiration line excursion (RLE) – the sum of absolute change for each successive pair of respiration samples – using a sliding window of three seconds over the 15 second EW. For respiration rates in the normal range (10 to 22 cycles per minute) the sliding window would encompass ½ to 2 respiration cycles. The respiration measurement was the mean of all three second windows during the EW. For the 15 second EW at the 30 cps data rate, the feature extraction value was the means of the (15 – 3) * 30 = 360 three-second segments. Use of a sliding window in this manner means that feature extraction values are not dependent on the length of the EW and can be easily compared and optimized for different EW lengths – leading to potentially easier optimization of the EW. The response feature of interest is a reduction or suppression of respiration activity, that is expected to occur when a person attempts to conceal, or to avoid revealing or telegraphic, their deception. Although the automated feature extraction algorithm uses a dimensionless quantification of the RLE, human evaluators will observe respiration responses visually in plotted/displayed waveform patterns – as a subtle reduction of the respiration amplitude and as a subtle slowing of respiration rate.

EDA feature extraction. For EDA data, information was evaluated from stimulus onset to the end of the EW. Response peaks were identified as the change in EDA slope from positive (upward) to negative (downward) from 2.5 seconds after stimulus onset to the end of the EW. One additional response peak was also evaluated following the end of the EW if the EDA slope remained positive from 13.5 seconds to the first peak after the EW. This permits the extraction of information to the peak of response instead of the end of the somewhat arbitrary EW and also prevents the evaluation of a response peak after the EDA slope has turned negative late in the EW. Response onsets were identified as the onset of a positive slope segment (i.e., a change from negative or zero slope to positive slope) from a .5 second latency point (LP) to the end of the ROW. Use of the LP eliminates the need to evaluate the EDA slope prior to stimulus onset and ensures that responses that begin immediately with stimulus onset are not evaluated. When the EDA slope was already positive prior to stimulus onset and remained positive throughout the ROW a response onset was imputed statistically as a function of a statistically significant change in positive slope variance (with alpha = .001) for two adjacent one-second moving windows from the LP to the end of the ROW. This can be visualized by human experts as a substantial change in upward angle within a positive slope segment during the ROW. The extracted value was the maximum difference between a response peak and a preceding response onset. In simplistic terms the EDA response feature can be visualized as the maximum distance from a response onset during the ROW, after the LP, to a peak point during the EW.

Cardio feature extraction. Feature extraction for the cardio data was similar to that for EDA, but with two differences. Cardio data was extracted from a moving average of all recorded cardio data points. The moving average was calculated by passing the cardio data four times through a moving average filter of .5 seconds. The result of the moving average filter can be visualized as the mid-line between systolic and diastolic peaks. Inclusion of one addition response peak, after the 15 second EW, was retained if the EDA slope remained positive from 14.5 seconds to the first response peak after the EW. This change was needed improve the ability of the feature extraction procedure to tolerate the potential complexity of cardiovascular activity, which can sometimes be influenced by respiration activity in addition to cognitive, emotional and behavioral factors. Similar to the EDA feature, the cardio response feature can be visualized as the maximum distance from a response onset during the ROW to a peak point during the EW. One difference between the automated feature extraction and manual feature extraction is that human examiner will most often evaluate the cardio data at the diastolic baseline. This procedure is thought to improve the reliability of visual/manual feature extraction and is premised on a strong correlation between the information contained in the diastolic line and mid-line.

Numerical transformation and data reduction. All physiological responses were measured in dimensionless units – not intended to represent a physical quantity. This permits the scaling of data for visual display and plotting with no effect on the numerical transformations involving the comparison of RQ and CQ values. Data were assumed to be ordinal and intervalic. For EDA and cardio data greater extracted values were associated with greater changes in physiological activity. For respiration data the feature of interest to PDD testing is a reduction of respiration activity in respiration activity in response to RQs and CQs. This meant that smaller respiration values were associated with a greater change in physiological activity for the respiration scores.

Assumptions and constraints. Recorded physiological data and extracted values were not assumed to be linear, and no ratio assumptions were employed in the transformation of extracted response values to ESS integer values. However, some linear constraints were employed to prevent the extraction and scoring of extreme values. Extreme values were defined as less than 2% of the maximum scaling value for visual display and plotting. Values smaller than the 2% threshold were regarded as potential noise, and therefore less likely to be an authentic response to the test stimuli, and were not used in the transformation of extracted values to ESS scores. Leave-one-out *z*-scores were calculated for all extracted values for each recording sensor within each recorded test chart. For EDA and cardio data *z*-scores in excess of 10 (i.e., 10 standard deviations) were regarded as data artifacts, possibly resulting from physical movement, and were excluded from the analysis. None of the responses exceeeded this value ($z > 10$) for this sampling data, though other samples have included response artifacts in excess of 10 standard deviations. Extracted response values were assumed to be monotonic with changes in physiology. That is, greater changes in physiological activity were assumed to be associated with differences in the extracted numerical values.

ESS integer scores. ESS integer scores were assigned using a three position scale of signed values [+, 0, -] similar to the procedure for the three-position scoring method. One difference between ESS scores and three-position scores is that ESS scores are obtained using only the primary response feature whereas the three-position and seven-position methods permit the combined use of primary and secondary responses (National Center for Credibility Assessment, 2017).

ESS integer scores were assigned to each RQ after comparing after comparing the RQ with a paired CQ. RQ and CQ pairs were selected via automated algorithm using the procedure described by Nelson (2017c). FZCT cases in the sampling data consisted of three RQs – named R5, R7, and R10. For each recording sensor R5 is compared to either adjacent CQ (C4 preceding the RQ or C6 subsequent to the RQ) depending on which CQ has the greater change in physiological activity. Use of two CQs for R5 in this manner is thought to benefit innocent/truthful persons in that a change in physiological activity at the RQ will have to exceed that of two CQs before deceptive score can be assigned – also will also provide them with two opportunities to produce a change in physiological activity at a CQ that exceeds that of the RQ. R7 is compared only to the preceding CQ (C6), and R10 is also compared only with the preceding CQ. Field practices permit the rotation of some questions during the various iterations of the test question sequence; this is intended to distribute or balance effects related to the position of each question in the sequence and also to dissuade examinees from memorizing or habituating to the question sequence. Regardless of the rotation of questions, the first RQ in the sequence is compared to the first two CQs, while the second and third RQs are compared only to the preceding CQs. Each paired RQ and CQ is sometimes referred to by polygraph field examiners as an analysis spot.

Respiration constraints and numerical scores. For respiration data, ESS integer scores of + sign value, indicative of truth-telling, were assigned when a greater change in physiology was observed in response to the CQ, while scores of – sign value, indicative of deception, were assigned when a greater change in physiology was observed in response to the RQ. In contrast to the EDA and cardio data, greater changes physiology are observed in respiration data as smaller extracted values – indicative of a greater reduction or suppression of respiration activity. To prevent the analysis of extreme changes or extreme values that may result from voluntary or deliberate activity – such as that sometimes observed by persons attempting to alter or fake their test data and results – a maximum respiration constraint ratio of 1.5:1 was employed on the RQ and CQ analysis spots. This constraint is intended to prevent the assignment of a signed integer score when a takes a deep breath or holds their breath in response to an RQ or CQ. Additionally, a minimum respiration constraint ratio of 1.25:1 was used to prevent the assignment of numerical score to response differences that are not due to the test stimuli – and which may be considered noise resulting from either normal/uncontrolled variation in respiration activity, or due to observed instability that some persons exhibit in their respiration rate and amplitude.

ESS scores from the abdominal and thoracic respiration sensors are combined into a single ESS score using the procedure described by Nelson and Krapohl (2019). This procedure is common to other manual scoring methods and will be familiar to many field practitioners. Scores are combined to a value of zero (0) when the sign values are opposite for the thoracic and abdominal sensors, and are collapsed to a single singed score when they are not opposite.

EDA and cardio constraints and numerical scores. For EDA and cardio data, ESS integer scores of + sign value, indicative of truth-telling, were assigned when a greater change in physiological activity was observed at the CQ. Greater changes in physiology are observed in EDA and cardio data as larger extracted values. Scores of – sign value, indicative of deception, were assigned when a greater change in physiological activity was observed in response to the RQ. Scores of 0 (0 sign value) were assigned when there was no observable or appreciable difference between the responses to an RQ and CQ analysis spot. A minimum constraint was used to prevent the assignment of score to EDA and cardio for RQ and CQ analysis spots for which the observed difference in response magnitude was small. The constraint selected for this project was a ratio of 1.05:1, for RQ and CQ analysis spots. This constraint was the result of step-wise optimization of correlation and receiver operating characteristic (ROC) coefficients with other data. Differences smaller than 5% are more likely to be the result of physiological noise, and may also be the result of unknown influence on the EDA data when using an auto-centering EDA solution for which the design characteristics are unknown or undocumented.

Weighted EDA scores. ESS integer scores for EDA data are weighted more than for other recording sensors. This is accomplished by doubling all + and – integer values to +2 and -2. The effect of this is to approximate the structural and statistical coefficients that have been reported in numerous studies on PDD data analysis and computer algorithm development over a period of nearly five decades.

[Refer to Nelson (2019) for a literature survey on structural and statistical coefficients for respiration, EDA, cardio and vasomotor data.]

Data reduction. Because ESS scores are signed integer values, data reduction is a simple matter of summation. Subtotal scores are summed for each RQ, including all sensors and all iterations of the question sequence. Subtotal scores are then summed for a grand total score. Although field practices and other analyses may often make use of subtotal scores, this project involves the analysis of only the grand total score.

Multinomial likelihood function and numerical cutscores

Likelihood function. The simplest form of likelihood function is a numerical cutscore that correspond to a known empirical likelihood of a correct or incorrect classification. Formally, a likelihood function is a tool – including possibly a mathematical or statistical formula, computer function or published reference table – that can be used to calculate or obtain a statistical value for the observed test data. Cutscores for ESS scores can be obtained from multinomial reference tables and Bayesian analysis.

Both grand total cutscores and subtotal cutscores can be used to classify the test data as either indicative of deception or truth-telling – the contextual allegory of the more general terms positive and negative. Published studies have consistently shown that grand total scores provide the highest rates of classification accuracy. [Refer to APA (2011) for a summary of effect sizes for validated polygraph techniques.] This is not be surprising when considering that grand total scores make use of more information than the question subtotal scores and will therefore provide reduced variation and more opportunity for data to converge. This sometimes referred to as the weak law of large numbers (Dekking, 2005) – a related to the central limit theorem which states that the means of randomly selected samples will be normally distributed and will converge towards the unknown population mean. It is the main reason that it is advantageous to have many samples (for which meta-analysis can also be used) and the reason that larger samples are preferred over smaller samples.

The question of great importance is this question: what numerical cutscores are most effective or most efficient to classify test results as indicative of deception or truth-telling? Or, more precisely, what probabilistic inferences about deception and truth-telling can be made about the numerical cutscores and resulting classifications? Because scientific testing and scientific test data analysis is inherently probabilistic (given that the purpose of any scientific test is to quantify a phenomena of interest that cannot be subject to physical measurement), field examiners and program managers will be primarily interested in this more practical version of the same question: what numerical cutscores will provide an optimal experience of correct vs incorrect outcomes? In scientific terms this is the question of selecting numerical cutscores that will optimize the desired observation of TN, TP, FN, and FP results. In the polygraph context the answer to this question will be considered with regard to the additional outcome potential for inconclusive outcomes.

Analytic theory of polygraph testing. The analytic theory of polygraph testing – under which the multinomial distributions of ESS and three-position scores are calculated – holds that greater changes in physiology will be loaded at different types of test stimuli (i.e., relevant and comparison questions) as a function of deception or truth-telling in response to the relevant or target stimuli. [Refer to Nelson (2016) for a discussion of the analytic theory of the polygraph test.] In polygraph testing, some uncontrolled variation is expected at the level of each sensor and each presentation of each RQ (e.g., it is not expected that scores will be of uniform sign value). To the degree that the theory of PDD testing is valid (supported by evidence), and PDD sensors record valid data (data and scores that are loaded as a function of deception or truth-telling and not mere randomness) the convergence of subtotal and grand total scores can be used to make statistical inferences about reality – the degree to which a person is probably deceptive or probably truthful. In other words, it is the aggregation of subtotal and grand total scores that will be used to classify the test data as indicative of deception or indicative of truth-telling.

Aggregation of scores from multiple RQ, multiple CQs, multiple recording sensors, and multiple iterations of the question relies on the law-of-large-numbers (LLN) – which for the aggregation of PDD scores will converge to become loaded to a value of either + or – sign value as a function of deception or truth-telling in response to the RQs. The LLN also provides insight as to why overall classification accuracy with grand total scores is expected to continue to outperform overall classification accuracy with subtotal scores.

Multinomial distribution is calculated under the analytic theory. The mathematical/ statistical distribution of data values (i.e., all possible ESS scores and the probabilities associated with each) can be characterized empirically, by obtaining data from reality. A distribution of ESS scores can also be calculated using only information subject to mathematical and logical proof under a proposed theory. Mathematical characterization of a distribution of scores is often accomplished under the null-hypothesis to a theory. This is because it is often difficult (read: impossible) to mathematically characterize a proposed theory while the (null-hypothesis) can often be easily characterized as a distribution of random values. A well-known distribution is the Gaussian or normal (bell-curved) distribution. We use our mathematical knowledge of statistical distribution to make inferences about individual cases relative to the population of all possibilities that is represented by the statistical distribution. In the polygraph context, because there is a finite, though large, number of all possible combinations of ESS scores for all iterations of all questions and all sensors, the statistical distribution of ESS scores is not Gaussian, but is multinomial. The distribution of ESS scores is multinomial because there are three possible values for each score.

The multinomial distribution of ESS scores will exhibit a bell shape, somewhat similar to the normal distribution, though with discrete values for each possible test score. Under the null-hypotheses – that scores are not loaded in any systematic way and can therefore be characterized as random – most multinomial scores will occur near the middle of the distribution (near zero) with only one possible way to achieve the maximum or minimum score (uniform + or – scores for every iteration of every question and every sensor). There is a finite, though large, number of possible combinations of [+, 0, -] scores for each exam. There is also a finite number of ways to arrange the [+, 0, -] scores to achieve each possible score.

The multinomial distribution of scores is a list of all possible scores and the probability associated with each; it can be calculated using a combinatoric formula. It can also be calculated (sometimes more easily and quickly) via Monte-Carlo simulation. (Multinomial calculations during the Manhattan Project were an impetus for the development of Monte Carlo statistical methods.) The multinomial distributions for ESS and three-position scores (Nelson, 2017a; 2018a) are an exact calculation. Most importantly, our knowledge and information about the multinomial distribution of ESS scores can be used to make statistical inferences about reality (i.e., classifications under uncertainty). All that is necessary is to first calculate the likelihood statistic for an observed score, if loaded for deception or truth-telling, and then use the statistical value from the multinomial distribution as a likelihood function in Bayesian analysis of the likelihood of deception or truth-telling.

In addition to ESS scores, a multinomial distributions have also been published for three-position scores (Nelson, 2018a). This is possible because the three-position method relies on the bigger-is-better rule for which reactions that are recorded and measured, regardless of whether using standardized or dimensionless/arbitrary measurement units, are objectively either larger, smaller or equivalent. These differences are larger because there is mathematical proof that successive numbers, whether positive or negative, can in factual reality, represent in larger and smaller quantities – including when those quantities are not assigned a standardized measurement unit. For this reason, results in this analysis were also calculated for grand total scores of Federal three position scores. Unfortunately, no multinomial distribution exists for Federal seven-position scores – due to arbitrary decisions (i.e., without mathematical proof) as to the differences in physiological activity that correspond to the seven-position scale values. Automation of seven-position scores cannot be accomplished using only facts and information subject to mathematical and logical proof, and for this reason questions about classification accuracy of seven-position grand total score were not addressed in this project.

Bayesian multinomial cutscores. Something that would be of great convenience would be to determine numerical cutscores that provide both a statistical classifier and also provide information about the practical meaning of the probabilistic strength of the classification. Multinomial cutscores for ESS cores (and three-position scores) together with Bayesian analysis do just that. Whereas early work on polygraph algorithms relied on statistical classifiers that were not intended to offer practical intuition or practical inference, multinomial cutscores, calculated using Bayesian methods, quantify both the practical or systematic likelihoods associated with deception or truth-telling in addition to the random error estimate associated with different outcomes that may result from the analysis of other data not available to the present analysis. Bayesian analysis is based on an assumption that the available sample/test data are all of the information available with which to make a conclusion (Stone, 2013; Winkler, 1972). In contrast, frequentist inference is based on an assumption that the available sample/test data are informative of the other data and information that could potentially be obtained from the universe and reality as it pertains to the individual and the behavioral target of a PDD investigation. [See Nelson (2017d) for a brief description of Bayesian analysis and null-hypothesis significance testing.] It is often the case the scientists and scientific methods may utilize a combination of frequentist and Bayesian assumptions. [Refer to Nelson (2018c) for a description of Bayesian analysis and the ESS-M.]

ESS Multinomial cutscores for grand total scores. For grand total scores with FZCT sample cases the multinomial grand total cutscores are - 3 or lower from deceptive classifications and +3 or greater for truthful classifications. There cutscores were selected from the multinomial distribution of all possible ESS scores (also the distribution of all possible ESS cutscores) at the point for which the random error estimate – indicated by the lower-limit of the Bayesian credible interval – provides a statistically significant likelihood (with alpha = .05) of continuing to observe the same analytic result, despite expected variation, if it were possible to repeat the examination or analysis numerous times. [Refer to Nelson (2018d) for a graphical illustration on the calculation of Bayesian ESS-M cutscores.]

Multinomial cutscores for three-position grand total scores. For three-position scores the multinomial cutscores can be calculated using the same Bayesian analytic methods as for the ESS. Multinomial cutscores for grand total scores of three-position scores are -2 or lower for deceptive classifications and +2 or greater for truthful classifications. [Also refer to Nelson (2020) for a tabular demonstration of ESS-M and three-position cutscores for a range of prior probabilities and different alpha levels for deceptive and truthful classifications.] Table 1 shows the multinomial cutscores for grand total scores with the three-position scoring method, along with the traditional cutscores for grand total scores.

Traditional cutscores. Traditional numerical cutscores were selected initially for older and more complex seven-position scoring methods; they too have been initially derived empirically and heuristically, and then subject to subsequent analysis for their classification efficiency. Traditional cutscores for grand total scores for FZCT exams are -6 or lower for deceptive classifications and +6 or greater for truthful classifications. An important consideration

Table 1. Traditional and multinomial cutscores for grand total scores with ESS and Federal 3-position scores.

	Traditional		Multinomial	
	Deception Indicated	No Deception Indicated	Deception Indicated	No Deception Indicated
ESS	-6	+6	-3	+3
Three-position	-6	+6	-2	+2

here is that field practice standards for Federal examiners who use the FZCT do not involve the use of grand total scores alone, and will instead involve a combination of grand total and subtotal scores. Although it is tempting to delve here and now into an empirical investigation of those procedures, and although little work has been published on the topic of decision rules since Senter and Dollins (2003), the purpose of this project was only to advance the available knowledge on effect sizes for numerical cutscores for grand total scores.

Another important consideration is that traditional grand total cutscores are also used with the three-position scoring method, leading to higher rates of inconclusive results for the three-position scoring method (APA, 2011) and the need to devote additional resources toward the resolution of these. Higher inconclusive rates also create a context for the emergence or reliance on covert solutions to reduce their occurrence. Most importantly, traditional cutscores were first suggested decades ago for the earlier and more complex seven-position scoring methods, and have remained unchanged despite scientific innovations in PDD data analysis and despite known and expected differences in the distribution of possible scores. Continued use of these traditional cutscores is a reflection of the fact that, although perhaps sub-optimal, outcome effects are reasonably known, and a more optimal solution, ideally supported by both theory and scientific evidence, has not yet been decided upon.

Interpretation and classification of analytic results.

Interpretation, in this usage, refers to the translation of numerical and statistical test results into categorical test results for which consistent and rational actionable decisions can be made. Interpretation and classification of test result is accomplished procedurally through the use of structured decision rules. Because this project involves the study of grand total cutscores, the decision rule of interest is the GTR.

Grand-total-rule. Execution of the GTR is a matter of summing the subtotal scores to obtain a grand total score. The grand total score is then compared to the numerical cutscores for grand total scores. Multinomial cutscores for the ESS and three-position methods are shown in Table 3. For ESS the multinomial cutscores are -3 or lower for deceptive classifications, and +3 or greater for truthful classifications. For three-position scores the multinomial cutscores are -2 or lower for deceptive classifications, and +2 or greater for truthful classifications. These cutscores are assuming a prior probability of 0.5. Traditional numerical cutscores for grand total scores are -6 or lower for deceptive classifications, and +6 or greater for truthful classifications. Traditional cutscores were derived for early seven-position scoring methods. Intuition suggests they may be inefficient for ESS and three-position scores – leading to higher rates of inconclusive results and over-reliance on subtotal scores. Although the use of subtotal scores, in addition to grand total scores, may improve classification with deceptive cases, this will introduce statistical multiplicity effects and may bias overall accuracy in unfortunate or unintended ways. For this reason, understanding and selection of optimal grand total cutscores may increase the accuracy effect sizes for the FZCT cases.

Bayesian analytic classification of deception or truth-telling. Multinomial grand total cutscores, for both ESS and three-position scores, provide a Bayesian posterior odds (systematic error) estimate of approximately 2:1 deception and truth-telling, permitting a 1-alpha x 100% = 95% likelihood of observing another analytic result of at least this value. In practical terms, test scores at this level are sufficient accept the notion that recorded physiological activity is loaded systematically, and to reject the notion that the scores are loaded in a random or meaningless un-interpretable/ un-classifiable way. Although 2:1 odds may not be spectacular, it is important to recognize that classifications made at a score of +/- 2 or +/- 3 cannot, when considering the range of the distribution of possible scores, be reasonably expected to provide spectacular accuracy. Equally important, posterior odds of 2:1 may provide actionable knowledge for some circumstances. For example: consider the information that the odds of a particular bridge collapsing under weight are estimated at 2:1. Many reasonable persons might be quite hesitant to make use of that bridge. Of course, circumstances will also exist that may require a stronger basis of actionable probabilistic information than 2:1 posterior odds. For most estimated prior probabilities, these needs can be met via the multinomial reference data and the selection of numerical cutscores that will constrain systematic and random error rates to required levels.

Results

Classifications for deception and truth-telling were calculated for each of the two FZCT samples. Because polygraph field practitioners commonly discuss test accuracy effects in terms of the proportions of correct, incorrect and inconclusive conclusions, accuracy effects are presented in this way – instead of using effects sizes that compare classifications to chance levels.

Results for sample 1, n=100 confirmed FZCT field exams.

There were no cases that changed from positive to negative classification and no cases that changed from negative to positive classification as a result of the scoring method or type of cutscore for this sample (n=100) of confirmed field cases. Table 2 shows the test

Table 2. Grand total classifications for n=100 FZCT field sample with ESS and three-position scores

	ESS scores		Three-position scores	
	Traditional cutscores	Multinomial cutscores	Traditional cutscores	Multinomial cutscores
Error	[5] .05 (.02) {.01 to .10}	[5] .05 (.02) {.01 to .10}	[3] .03 (.02) {.01 to .07}	[5] .05 (.02) {.01 to .10}
Inconclusive	[35] .35 (.08) {.15 to .46}	[11] .11 (.05) {.01 to .18}	[53] .53 (.09) {.25 to .58}	[14] .14 (.05) {.01 to .21}
Correct	[60] .92 (.03) {.85 to .98}	[84] .94 (.02) {.89 to .99}	[44] .94 (.04) {.85 to .99}	[81] .94 (.03) {.88 to .99}
Sensitivity (TP)	[33] .66 (.07) {.53 to .79}	[44] .88 (.05) {.78 to .96}	[28] .56 (.07) {.43 to .70}	[43] .86 (.05) {.75 to .95}
Specificity (TN)	[27] .54 (.07) {.40 to .68}	[40] .80 (.06) {.69 to .91}	[16] .32 (.07) {.19 to .46}	[38] .76 (.06) {.64 to .87}
FN errors	[2] .04 (.03) {.01 to .10}	[2] .04 (.03) {.01 to .10}	[1] .02 (.02) {.01 to .07}	[2] .04 (.03) {.01 to .10}
FP errors	[3] .06 (.03) {.01 to .13}	[3] .06 (.03) {.01 to .13}	[2] .04 (.03) {.01 to .10}	[3] .06 (.03) {.01 to .13}
Inc guilty cases	[15] .30 (.06) {.18 to .43}	[4] .08 (.04) {.02 to .16}	[21] .42 (.07) {.29 to .55}	[5] .10 (.04) {.02 to .19}
Inc innocent cases	[20] .40 (.07) {.27 to .54}	[7] .14 (.05) {.05 to .24}	[32] .64 (.07) {.50 to .78}	[9] .18 (.05) {.08 to .29}
PPV	.92 (.05) {.82 to .99}	.94 (.04) {.86 to .99}	.93 (.05) {.83 to .99}	.93 (.04) {.86 to .99}
NPV	.93 (.05) {.83 to .99}	.95 (.03) {.88 to .99}	.94 (.06) {.80 to .99}	.95 (.03) {.88 to .99}
Correct guilty cases	.94 (.04) {.86 to .99}	.96 (.03) {.89 to .99}	.97 (.03) {.89 to .99}	.96 (.03) {.88 to .99}
Correct innocent cases	.90 (.06) {.78 to .99}	.93 (.04) {.85 to .99}	.89 (.08) {.72 to .99}	.93 (.04) {.84 to .99}
Unweighted inc.	.35 (.05) {.26 to .44}	.11 (.03) {.05 to .18}	.53 (.05) {.43 to .62}	.14 (.03) {.07 to .21}
Unweighted accuracy	.92 (.03) {.85 to .98}	.94 (.02) {.89 to .99}	.93 (.04) {.84 to .99}	.94 (.03) {.88 to .99}

* Cells show the [frequency] in addition to the bootstrap estimate of the mean, (standard deviation) and {95% confidence interval}.

accuracy metrics for classifications using the GTR with both traditional cutscores and multinomial cutscores for ESS and three-position scores. Included in Table 2 are the error and inconclusive rates, along with the proportion of correct classifications excluding inconclusive results. Also included in Table 2 are the sensitivity (TP) and specificity (TN) rates, along with FN and FP error rates. Other metrics in Table 2 are the positive predictive value (PPV) calculated as the proportion of true positive results and all positive results, and negative predictive value (NPV) which is the proportion of true negative and all negative classifications. Also shown are the proportion of correct decisions for guilty and innocent cases excluding inconclusive result, along with the unweighted accuracy and unweighted inconclusive rates.

Inspection of the rows in Table 2 indicates that the confidence intervals are substantially overlapping for the proportions of errors produced by the four treatments. However, some differences can be observed in sensitivity, specificity and inconclusive results. Both sensitivity to deception and specificity to truth-telling were greater for ESS scores and for multinomial cutscores. Inconclusive rates were lower for ESS scores and for multinomial cutscores. The frequency of TP and TN results was greater for the ESS and multinomial cutscores and lower for three-position and traditional cutscores. The frequencies of inconclusive results were higher for traditional cutscores and lower for multinomial cutscores.

Results for sample 2, n=60 confirmed FZCT field exams.

For the second sample of n=60 FZCT cases, there were no cases for which the classification changed from positive to negative or from negative to positive as a result of the scoring method or cutscore type.

Inspection of the rows in Table 3 indicates that results with the second FZCT sample paralleled those of the first sample. Confidence intervals are substantially overlapping for the accuracy metrics for correct classifications. Sensitivity to deception and specificity to truth-telling were greater for ESS scores and for multinomial cutscores. Inconclusive rates were lower for ESS scores and for multinomial cutscores.

Table 3 shows the same test accuracy metrics for ESS and Federal three position scores for the second archival sample.

Table 3. Grand total classifications for n=60 FZCT field sample with ESS and three-position scores[*]

	ESS scores		Three-position scores	
	Traditional cutscores	Multinomial cutscores	Traditional cutscores	Multinomial cutscores
Error	[2] .03 (.02) {.01 to .08}	[2] .03 (.02) {.01 to .08}	[1] .02 (.02) {.01 to .05}	[5] .08 (.04) {.02 to .15}
Inconclusive	[21] .35 (.10) {.10 to .50}	[7] .12 (.04) {.01 to .13}	[32] .53 (.11) {.22 to .63}	[6] .10 (.06) {.01 to .2}
Correct	[37] .95 (.04) {.87 to .99}	[51] .96 (.03) {.91 to .99}	[27] .96 (.04) {.89 to .99}	[49] .91 (.04) {.82 to .98}
Sensitivity (TP)	[21] .68 (.09) {.50 to .84}	[29] .94 (.05) {.83 to .99}	[17] .55 (.09) {.37 to .73}	[27] .90 (.05) {.79 to .99}
Specificity (TN)	[16] .53 (.09) {.34 to .72}	[22] .73 (.08) {.57 to .89}	[10] .33 (.09) {.17 to .50}	[22] .73 (.08) {.56 to .88}
FN errors	[0] .03 (.03) {.01 to .11}	[0] .03 (.03) {.01 to .11}	[0] .03 (.03) {.01 to .11}	[1] .03 (.03) {.01 to .11}
FP errors	[2] .07 (.05) {.01 to .17}	[2] .07 (.05) {.01 to .17}	[1] .03 (.03) {.01 to .11}	[4] .13 (.06) {.03 to .26}
Inc guilty cases	[9] .29 (.08) {.14 to .46}	[1] .03 (.03) {.01 to .11}	[13] .42 (.09) {.24 to .60}	[2] .07 (.05) {.01 to .17}
Inc innocent cases	[12] .40 (.09) {.23 to .59}	[6] .20 (.07) {.07 to .35}	[19] .64 (.09) {.45 to .81}	[4] .14 (.06) {.03 to .27}
PPV	.91 (.06) {.77 to .99}	.94 (.04) {.83 to .99}	.95 (.06) {.82 to .99}	.87 (.06) {.74 to .97}
NPV	.94 (.06) {.80 to .99}	.96 (.04) {.85 to .99}	.91 (.09) {.71 to .99}	.96 (.04) {.86 to .99}
Correct guilty cases	.96 (.05) {.85 to .99}	.97 (.03) {.88 to .99}	.95 (.05) {.81 to .99}	.96 (.03) {.88 to .99}
Correct innocent cases	.89 (.08) {.71 to .99}	.92 (.06) {.79 to .99}	.91 (.09) {.70 to .99}	.85 (.07) {.70 to .97}
Unweighted inc.	.35 (.06) {.23 to .47}	.12 (.04) {.05 to .20}	.53 (.06) {.40 to .65}	.10 (.04) {.03 to .18}
Unweighted accuracy	.92 (.05) {.82 to .99}	.94 (.03) {.87 to .99}	.93 (.05) {.81 to .99}	.91 (.04) {.82 to .98}

[*] Cells show the [frequency] in addition to the bootstrap estimate of the mean, (standard deviation) and {95% confidence interval}.

Analysis of the combined sample data

A two-way repeat measures ANOVA (scoring method x cutscore type) showed a significant interaction for inconclusive results [F (1,636) = 329.671, (p < .001)], indicating that observed differences in inconclusive rates for multinomial and traditional cutscores were different for ESS and three position scores. One way ANOVAs showed that the reduction of inconclusive results was statistically significant at the a=.05 level for ESS scores [F (1,318) = 4, (p = .046)], and also for the three position scores [F (1,318) = 12.308, (p < .001)].

A three-way repeat measures ANOVA for correct positive and negative classifications showed a significant three-way interaction, for criterion state (guilty vs innocence), scoring method (ESS, three-position) and cutscore type (traditional, multinomial) [F (1,632) = 54.282, (p < .001)]. Table 4 shows the ANOVA summary. All of the main effects and two-way interactions were also significant in the three-way ANOVA but were not interpretable due to the significant interaction effects.

Because differences in accuracy effects as function of cutscore type were the main interest for this project, a series of one-way contrasts was completed. For ESS scores the one-way effect was statistically significant for increased test sensitivity to deception [F (1,158) = 7.533, (p = .007)] and for increased test specificity to truth-telling [F (1,158) = 5.347, (p = .022)]. For three-position scores the one-way effect was also statistically significant for both increased test sensitivity to deception [F (1,158) = 11.148, (p = .001)] and test specificity to truth-telling [F (1,158) = 17.663, (p < .001)].

Risk ratios

To more adequately understand the differences in sensitivity, specificity and inconclusive rates shown by these data, risk-ratios were calculated after transforming the observed proportions to odds. Risk ratios are based on an assumption that observed proportions are an estimate of the likelihood or strength of the possibility of observing a similar outcome with any randomly selected member of the population, and are calculated as the ratio of the proportions observed for two different methods. In this project the comparison of interest is the risk-ratio for differences in outcomes for traditional and multinomial cutscores. Table 4 shows the risk-ratios for true-positive, true-negative and inconclusive results.

Risk ratios are informative as to the practical likelihood of differences in observed outcomes as they may be experienced for an individual or groups of cases. Risk ratios in Table 4 suggest that the use of multinomial cutscores with ESS scores may reduce the likelihood or occurrence of inconclusive outcomes to 34% of what would be expected from traditional cutscores. For three-position scores the risk of inconclusive outcomes may be reduced to approximately 20% of the risk of inconclusive results using traditional cutscores. However, actual rates in field practice the observed difference may not achieve these estimated differences because field practitioners may already engage in a variety of activities to resolve or reduce the occurrence of inconclusive results. Information shown in Table 4 indicates that guilty persons are 1.4 times more likely to be detected using multinomial cutscores, while innocent persons may be 1.6 times more likely to be classified as truthful.

Table 4. Risk-ratios for TP, TN, and inconclusive results for traditional and multinomial cutscore

	ESS scores	Three-position scores
Inconclusive	.34 (2.9)	.19 (5.3)
Sensitivity (TP)	1.38	1.38
Specificity (TN)	1.64	2.21

Conclusion

This project, concerned with the study of grand total cutscores, involved the use of the GTR with ESS and three-position scores using both traditional and multinomial cutscores. The GTR, although perhaps the simplest of all PDD decision rules, has been shown to provide the highest rates of overall classification accuracy among the variety of PDD decision rules. Accuracy effects were compared for ESS and three-position scores of event-specific polygraph exams using traditional numerical cutscores and multinomial cutscores for grand total scores. Although skill development at manual scoring with the ESS – as with PDD testing in general – is most effectively acquired as a function of both didactic or academic knowledge of standardized procedures and practical supervision and guidance under other experienced professionals, the basic concepts of the ESS are simple and highly structured, leading to the potential for an automated process that closely approximates the activities of human experts. During this project, to ensure that observed variance can be attributed to differences in numerical cutscores, and not due to expected variation within the inter-rater reliability limits of manual scoring methods, ESS and three-position scores and results were obtained via computer algorithm, including feature extraction, selection of RQ and CQ analysis spots, numerical transformation, data reduction application of the GTR with both multinomial and traditional cutscores.

Data for two different archival samples of confirmed FZCT field exams were used. These archival samples have been characterized by human scorers as challenging, though previously reported effect sizes for both human and computer algorithms is consistent with those shown herein. Results are shown separately in table form for each of the samples. Data for two samples were then combined for statistical analysis of potential differences in effect sizes for ESS and three-position scores. For the combined samples, significant differences were observed for test sensitivity to deception, test specificity to truth-telling, and the proportion of inconclusive results. Use of multinomial cutscores reduced the occurrence of inconclusive results and increased both sensitivity to deception and specificity to truth-telling. Use of archival data permits the direct comparison of observed effect sizes with previously reported effect using the same data with other scoring methods for which field examiners have intuitive knowledge and experience of their effectiveness.

Of particular interest in this project is that none of the deceptive or truthful classifications were reversed as a result of the selection of traditional or multinomial cutscores for grand total cutscores with either the ESS or three-position scores. Another interesting observation in this project is that accuracy effects for correct classifications, including PPV, NPV, and the proportion of correct classifications excluding inconclusive results within the guilty and innocent, along with the unweighted accuracy excluding inconclusive results were substantially similar for both traditional and multinomial cutscore with both ESS and three-position numerical scores. However, use of multinomial cutscores increased sensitivity to deception by factor of 1.4 for both ESS and three-position scores, while increasing specificity to truth-telling by a factor of 1.6 for ESS scores and by a factor of 2.2 for three-position scores. Multinomial cutscores reduced the occurrence of inconclusive results by a factor of 2.9 for ESS scores, and by a factor of 5.3 for three-position scores. Use of multinomial cutscores and grand total scores with the FZCT format, a three-question single issue format, achieves the level of accuracy requires by the American Polygraph Association Standards of Practice for evidentiary exams – those exams conducted with an expectation that the results and information may be used as information in a legal proceeding (American Polygraph Association, 2018)) – for both ESS and three-position scores.

Potential practical implications of these results include the possibility of increasing the effectiveness of field polygraphs in correctly classifying both deception and truth-telling. Another practical implication is that it is a reasonable consideration, in terms of classification accuracy, for polygraph programs to make use of traditional numerical cutscores with ESS scores if the prospect of changing both score type (ESS or three-position scores) and cutscores presents an uncomfortable number of degrees of freedom for policy makers. Indeed, there is practical and scientific wisdom in changing one variable at a time while observing and gaining experience with different methods. More broadly, these results support that it is a reasonable consideration for field examiners and/or policy makers to consider using only the grand total score for the FZCT format – a finding for which there is no basis of information or theoretical rationale to suggest that it can be generalized to all single issue polygraph formats.

Like all projects, this one is subject to some limitations. One difference between the procedures used during this project and those used by field practitioners when manual scoring is that this project is limited to the analysis of grand total scores. Human experts in field practice may rely on different decision rules depending on agency policy. Although not all agencies will choose to make use of only the grand total score, grand total scores have been shown consistently in published studies to provide the highest overall rates of classification accuracy. Use of grand total scores in this manner is thought to provide the greatest insight into the influence of grand total cutscores on overall test accuracy regardless of the decision rules used in field practice.

This project did not attempt to study effects with decision rules other than the GTR. Study of interaction effects involving both numerical cutscores and decision rules that may make use of both grand total and subtotal scores would require a multivariate analysis that is beyond the scope of this analysis – intended to be a simple an intuitive descriptive survey of accuracy effects with grand total scores. A more comprehensive project would have investigated both the type of cutscores and the decision rule. However, such a project would expand the complexity of the analysis considerably, along with a corresponding increase in the complexity of the analysis and information from the analysis. Limiting this project to 2 dimensions – scoring type and cutscore type – was thought to provide information of potentially practical use while also addressing the analytic questions in some degree of depth.

There is reason to expect that some interaction exists between the decision rule and the selection of numerical cutscores. One obvious implication of these analytic results is that traditional numerical cutscores for grand total scores, though not inaccurate, appear to be inefficient. A consequence of this is that polygraph field practitioners, in addition to polygraph trainers, quality control personnel and program managers, may have come to rely more heavily than is ideal on subtotal score to remediate the inefficiency. A consequence of over-reliance on subtotal scores to remediate inefficiency is that use of subtotal scores will introduce statistical multiplicity effects than complicate the test accuracy effects – most often in ways that can make the test appear biased against innocent persons. Selection of a more optimal grand total cutscore may increase test accuracy with both guilty and innocent persons while potentially relieving some of the burden of multiplicity effects. The interaction of decision rules and numerical cutscores should be the topic of further analysis and study.

Another limitation of this study involves the three-position scores. In this project three position scores were achieved by flattening of EDA scores of the ESS scores. It is unknown to what degree these three-position scores may differ from those achieved in field practice contexts where examiners might make use of secondary response features and other semi-subjective practices that are not included in the ESS and which cannot be subject to automation. In principle, three position scores can be extracted using the exact same automated procedure as for ESS scores. Three-position scores can also be achieved using the more complex system of primary and secondary features that was developed for seven-position scores – and which is less easily amenable to automation. Regardless of this limitation, it is the view of the author that the three-position values herein are sufficiently representative for these results to provide some potentially useful information.

A final limitation is that inconclusive rates observed in this study, like all scientific studies, may be greater than those observed in field practice. This to be expected. Polygraph field examiners, and polygraph programs, are regarded as acting reasonable if they engage in efforts to resolve inconclusive results at the level of each individual case. In research of this type such efforts would amount to manipulating the research outcome. For this reason, no effort is made, in projects of this type, to resolve or reduce the occurrence of inconclusive results at the individual case level. Differences in inconclusive results are a reflection of the analysis method, and will necessarily be greater than inconclusive rates in field practice.

With consideration for the acknowledged limitations, accuracy effects observed in this analysis place the FZCT in the range required by the APA standards of practice for evidentiary examinations (APA, 2018). Evidentiary exams are those for which the test is conducted with the intention of introducing the test result as a basis of information in a legal proceeding. Accuracy rates observed herein equal or exceed those of other evidentiary PDD formats. Continued interest and continued research is recommended for ESS scores, the GTR and the use of multinomial cutscores.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. Polygraph, 40, 194-305. [Electronic version] Retrieved August 20, 2012, from http://www.polygraph. org/section/research-standards-apa-publications.

American Polygraph Association (2018). APA Standards of Practice (Effective September 1, 2018). Retrieved (January 20, 2019) from https://www.polygraph.org/apa-bylaws-and-standards.

Butterworth, S. (1930). On the theory of amplifiers. Experimental Wireless and the Wireless Engineer, 7, 536-541.

Dekking, Michel (2005). A Modern Introduction to Probability and Statistics. Springer.

Department of Defense (2006). Federal psychophysiological detection of deception examiner handbook. Available from the author. Reprinted in Polygraph, 40 (1), 2-66.

Editorial Staff. (2019). Introduction to the NCCA ASCII Standard. Polygraph and Forensic Credibility Assessment, 48(2), 125-135.Polygraph & Forensic Credibility Assessment,

Krapohl, D. J. (2002). Short report: Update for the objective scoring system. Polygraph, 31, 298-302.

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

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

National Center for Credibility Assessment (2017). Test Data Analysis: Numerical Evaluation Scoring System Pamphlet. Available from the author. (Retrieved from http://www.antipolygraph.org on 4-13-2019).

National Research Council (2003). The Polygraph and Lie Detection. Washington, D.C.: National Academy of Sciences.

Nelson, R. (2015). Scientific basis for polygraph testing. Polygraph 41(1), 21-61.

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

Nelson, R. (2017a). Multinomial reference distributions for the Empirical Scoring System. Polygraph and 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. Polygraph and Forensic Credibility Assessment, 46 (2), 116-131.

Nelson, R (2017c). Heuristic principles to select comparison and relevant question pairs when scoring any CQT format. APA Magazine, 50(1), 73-83.

Nelson, R. (2017d). Five-minute science lesson: Bayesian and frequentist statistics – what's the deal? APA Magazine, 50(4), 89-95.

Nelson, R. (2018a). Multinomial reference distributions for three-position scores of comparison question polygraph examinations. Polygraph and Forensic Credibility Assessment, 47(2), 158-175.

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

Nelson, R. (2018c). Five minute science lesson: Bayes' theorem and Bayesian analysis. APA Magazine, 51(5), 65-78.

Nelson, R. (2018d). Practical polygraph: a tutorial (with graphics) on posterior results and credible intervals using the ESS-M Bayesian classifier. APA Magazine, 51(4), 66-87.

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

Nelson, R. (2020). Multinomial cutscores for Bayesian analysis with ESS and three-position scores of comparison question polygraph tests. Polygraph and Forensic Credibility Assessment, 49(1), 61-72.

Nelson, R. & Krapohl, D. (2011). Criterion validity of the Empirical Scoring System with experienced examiners: Comparison with the seven-position evidentiary model using the Federal Zone Comparison Technique. Polygraph, 40, 79-85.

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. Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B., & Oelrich, M. (2011). Using the Empirical Scoring System. Polygraph, 40, 67-78.

Nelson, R. & Krapohl, D. K. (2017) Practical polygraph: a recommendation for combining the upper and lower respiration data for a single respiration score. APA Magazine. 50(6), 31-41.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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.

Stone, J. (2013). Bayes' Rule: A Tutorial Introduction to Bayesian Analysis. Sebtel Press.

Winkler, R. L. (1972). An Introduction to Bayesian Inference and Decision. Holt Mc-Dougal.