Brute-Force Comparison: A Monte Carlo Study of the Objective Scoring System version 3 (OSS-3) and Human Polygraph Scorers

Raymond Nelson[1], Donald J. Krapohl[2], and Mark Handler[3]

Abstract

The authors describe the Objective Scoring System, version 3 (OSS-3) scoring algorithm, and used brute-force statistical methods to compared its accuracy to previously described scoring algorithms and human examiners, using the OSS development sample (N=292) of confirmed single-issue field investigation polygraphs, and a second sample (N=100) of confirmed single-issue investigation cases. OSS-3 demonstrated balanced sensitivity and specificity and provided significant improvements over previous OSS versions in the form of reduced inconclusive results and increased sensitivity to deception. The improvement in specificity to truthfulness was not significant. The Gaussian-Gaussian decision model of OSS-3 was compared to a replication of an empirical Bayesian decision algorithm described by Kircher and Raskin (1988; 2002), and Raskin, Kircher, Honts, and Horowitz, (1988), that was trained on the OSS development sample using discriminate analysis. OSS-3 showed accuracy that met or exceeded that of the empirical Bayesian algorithm. Using Monte Carlo techniques and OSS-3 accuracy exceeded the average decision accuracy of the 10 human scorers, and 9 out of 10 individual scorers, on 6 dimensions of accuracy: overall decision accuracy, inconclusive results, sensitivity to deception, specificity to truthfulness, false negative, and false positive results. Interrater reliability for the 10 human scorers was evaluated using a double bootstrap of Fleiss' kappa, was consistent with previous reported reliability estimates (*k* = .59, 95% CI = .51 to .66), compared to the expected perfect reliability of the automated algorithm. A cohort of inexperienced polygraph examiner trainees was trained to evaluate the archival sample using a simplified set of scoring rules and optimized decision rules, intended to approximate the function of the new algorithm as reasonable as possible within human scorers. Decision accuracy for the trainees, using the simplified scoring instructions, was not statistically different from that for the experienced examiners. Interrater consistency for the inexperienced scorers was compared to the experienced scorers using a bootstrap resample of the 1000 iterations of the archival sample (N=100). Fleiss' kappa for the student examiner cohort was *k* = .61, which was not statistically different from the experienced scorers (*k* = .58). The computer algorithm can be expected to provide perfect reliability. The authors suggest that computer algorithms should be given more weight in quality assurance and field practices, though they caution the need for responsibility surrounding professional opinions and administrative decisions and policies. The authors also encourage further research into the possible development of a simplified rubric for polygraph hand-scoring.

Introduction

Comparison question polygraphy relies on the transformation of physiological reactions to mathematical representations that can be evaluated for empirical classification efficiency or statistical significance. Polygraph scoring research has

[1] Raymond Nelson is the developer of OSS-3 and past contributor to this journal. Requests for reprints should be directed to him at raymond@raymondnelson.us

[2] Donald Krapohl is Past President of the APA (2006-2007) and regular contributor to *Polygraph*.

[3] Mark Handler is a regular contributor to *Polygraph*.

Disclaimer: The OSS-3 algorithm described in this report was developed by the authors to provide an open source, objective and scientifically defensible method for analyzing polygraph data, and was offered openly to the polygraph community. None of the authors has a financial interest in OSS-3.

been almost entirely empirical, with less emphasis on statistical evaluation of the distributions of truthful and deceptive scores. The Objective Scoring system (OSS) (Krapohl & McManus, 1999; Krapohl, 2002) is an exception to this trend, and was designed and normed using the principles of statistical decision making and signal detection theory.

All scoring techniques for comparison question polygraphs involve the transformation of the observation or measurement of differential physiological reactivity to various question types to a numerical index representing the saliency of the target stimulus. In the case of polygraph hand-scoring techniques, data are commonly transformed to ordinal seven-position values, between -3 and +3 (Backster, 1963; 1990; Bell, Raskin, Honts, Kircher, 1999; Handler, 2006; Research Department Staff, 2006; Swinford, 1999) through observation of a test subject's differential reactivity to various test questions within each component sensor. Data are transformed for each presentation of each target stimulus, for each component sensor, and then aggregated to formulate a conclusion.

While the Krapohl and McManus (1999) system is intended to provide a uniform septile distribution of scores, other hand-scoring systems have not completely described the anticipated scoring distributions. There has been some investigation into the frequency of occurrence of numerical values and point assignments (Capps & Ansley, 1992; Krapohl, 1998). Three-position ordinal scales between -1 and +1 have been suggested (Capps & Ansley, 1992b; Van Herk, M., 1990), and investigated (Harwell, 2000; Krapohl, 1998). Blackwell (1998) concluded that seven-position scoring outperformed three-position scoring, but it should be viewed cautiously because there was no adjustment of cutscores for the differences in the distributions of three-position totals compared to seven-position totals.

Polygraph hand-scoring methods vary in their transformation methods. Numerical scoring, as taught at the Defense Academy for Credibility Assessment (Research Department Staff, 2006), requires evaluation of numerical ratios for seven-position score assignments, which imposes requirements for physical measurement of the data.

Some investigators have described rank-order analysis (Gordon, 1999; Gordon & Cochetti, 1987; Honts & Driscoll, 1988; Krapohl, Dutton, & Ryan, 2001; Miritello, 1999). Rank schemes are easily understood nonparametric methods and have shown some promise. However, because ranking replaces the natural variance of the data with a uniform rank variance, rank order schemes may not extend well to scoring systems intended to evaluate multiple simultaneous investigation targets. Efforts to apply rank order schemes to multi-facet or mixed issues examinations can be investigated empirically but will lack both face and construct validity under attempts to reconcile those methods with statistical theory involving the normal variance of differential reactivity to individual investigation targets. Miritello's (1999) description of a rank-order method for mixed question exams is lacking both normative data and a decision model.

Some hand-scoring systems include more features and rules than others. Systems developed by Kircher and Raskin (1988), as described by Bell et al. (1999) and Handler (2006) have systematically reduced interpretable features by excluding those that cannot be reliably and consistently measured, or are not supported by multiple studies. Other differences among the various scoring systems include the interpretation of pneumograph response data before or after the point of answer, interpretation of non-measured criteria such as complexity and changes in respiratory data, and the inclusion of arbitrary numerical data into measurement values when time-domain metrics are described in physical dimension instead of units of time. (see Kircher & Raskin, 2002, and Podlesny & Truslow, 1993, for more discussion regarding this concern.)

Scoring features of the Utah system (Bell, Raskin, Honts, & Kircher, 1999; Handler, 2006), are supported by complete description of their development and validation through discriminate analysis (Kircher & Raskin, 1988; 2002), and are similar to the features described in ASTM standard E-2229-02 (ASTM International 2002) and those currently taught at the Defense Academy for Credibility Assessment (Research Staff, 2006). While CPS (Kircher & Raskin, 1999) and OSS (Krapohl & McManus,

1999) employ features that are familiar to human scorers, PolyScore (Olsen, Harris, & Chiu, 1994; Harris & Olsen, 1994) uses features that were obtained through logistic regression and would be unfamiliar to human examiners. Other available algorithms include Chart Analysis and AXCON (Dollins, Krapohl & Dutton, 2000) in the Axciton computer polygraph (Axciton Systems, Houston TX) and Identifi (Dollins et al., 2000). Those methods employ features, decision models, and normative data that are not completely described in publication.

Despite their differences, polygraph hand-scoring systems are consistent in that greater saliency or differential reactivity to the investigation targets (relevant questions) are correlated with deception. In hand-scoring systems these reactions are assigned negative (-) integer values. Segments of greater differential reactivity to comparison stimuli indicate the investigation targets are less salient or correlated with truthfulness, and are assigned positive (+) integer values.

Hand-scoring values are summed within each test series, for each presentation of each target stimulus, and then summed between test series for each target stimulus. Finally, data for the several target stimuli are summed for a grand total. Field examiners refer to a pair of relevant and comparison stimuli as a *spot*, though the term also applies to the sum of repetitions of each separate investigation target. Polygraph hand-scoring decision policies utilize both total and spot scores, depending on whether the spot scoring rule is used (Capps & Ansley, 1992c). Several studies have investigated the contribution of the spot scoring rule. While previous OSS versions employed decision policies based on total scores, spot scoring has been the predominate method for decision policies pertaining to multiple facets of an alleged incident, or a mixed set of target issues with no known allegations. Although Krapohl and Stern (2003) used the terms *multiple-facet* and *multiple-issue*, the differences between these terms are not universally understood by many field examiners. This can lead to practical and mathematical problems. We will use the expressions *multiple-facet* and *mixed-issue* here to capture the importance of the dependence or independence of the target stimuli. (See Krapohl & Stern, 2003, for a description of multiple-issue testing and the use of combined testing strategies in medical and related testing contexts.)

Various strategies exist to maximize decision accuracy for both multiple-facet and mixed-issue polygraph examinations that are based on a straightforward adjustment of cutscores. Several studies have focused on spot scoring rules, that is, triggering a call of Deception Indicated (DI) based upon strong negative scores to a single relevant question. The use of spot score rules generally improves sensitivity to deception, though at a cost of increased false-positive errors (Senter, 2003; Senter & Dollins, 2002; Senter & Dollins, 2004; Senter, Dollins & Krapohl, 2004). A more careful examination of the underlying statistical distributions would have predicted this effect. It is hardly surprising that we have found no discussion regarding inflated alpha and corrective measures when completing multiple simultaneous significance tests, because polygraph hand-scoring research has generally not investigated the distributions of scores or statistical analysis. Senter's (2003) report that two-stage rules can improve the overall decision accuracy of MGQT exams and provide a more optimal balance of sensitivity and specificity was a procedural solution. It stopped short, however, of statistical procedures such as the Bonferonni correction to alpha, omnibus analysis through the use of ANOVA procedures, or statistical procedures such as the Tukey test, which are designed to manage the complications of multiple comparisons attending to the multiple-facet and mixed-issue examinations. OSS-3 uses Senter’s two-stage rules for ZCT and MGQT examinations along with statistical corrections that accommodate the morphology of the underlying distributions.

In our design and laboratory model for the OSS-3 algorithm, we included a Kruskal-Wallis test, as a nonparametric ANOVA, to serve as an omnibus assessment of the significance of differences between the target stimuli of mixed-issues examinations. The present study addresses only the use of OSS-3 with event-specific single-issue polygraphs Zone Comparison Techniques. Additional capabilities of the algorithm will be described in other studies. Design protocols for OSS-3 also include the use of a Test of Proportions, to monitor the distribution of artifacted and uninterpretable values during an examination (Menges, personal communication 3/12/2008). That portion of the OSS-3 algorithm was not evaluated in the present study.

Objective Scoring System, Versions 1 and 2

OSS (Krapohl & McManus, 1999) is based on three simple and mechanically repeatable measurement aspects of polygraph waveforms, called "Kircher features," (Dutton, 2000) which were first described by Kircher & Raskin, (1988). The three Kircher features include: Respiration Line Length (Timm 1982; Krapohl & Dutton, 2001), electrodermal phasic amplitude of increase, and cardiovascular phasic amplitude of increase. Harris, Horner, and McQuarrie (2000) recommended these same physiological indicators as the most robust feature set, and reported these three features as capable of replicating the 7-position numerical scoring system that was in use at the Department of Defense at that time. Kircher, Kristjansson, Gardner, and Webb (2005), provided further argument for this simple set of features as the most robust and reliable feature set for present-day polygraph scoring. These features provide desirable attributes, including that they are easily understood by human examiners or reviewers, are similar to features used in hand-scoring, and they can be mechanically measured with perfect reliability. Both OSS and the Computerized Polygraph System (CPS) (Kircher & Raskin, 1999; 2002) are based on the three Kircher features.

Krapohl (1999) suggested the use of a dimensionless R/C ratio transformation of the Kircher features which became the foundation of earlier OSS versions and was retained in OSS-3. Ratio transformation reduces the mathematical comparison of values to a single ratio value for each measured component sensor, for each presentation of the target stimuli. These ratios are dimensionless in that the physical units of measurement are canceled out algebraically during calculation. R/C ratios are also asymmetrical, in that the distribution of all possible R/C ratios will be a positively skewed distribution of lognormal shape, consisting of all positive real numbers, with a mean of one, an infinite number of possible values between zero and one, and a similarly infinite number of values between one and infinity.

OSS procedures involve the calculation of a physically dimensionless ratio of differential reactivity to various question types, and the transformation of those ratios to a uniform septile distribution of integer values from -3 to +3. OSS total scores are then summed and subject to a Gaussian signal detection model (Wickens, 1991; 2002) that was described by Barland (1985). Krapohl and McManus (1999) provided tables of statistical significance that were constructed using normative data from a large sample of event-specific single-issue investigation polygraphs using Zone Comparison Techniques (ZCT) (Light, 1999) that included three relevant questions concerning a single target allegation, along with three comparison questions and three test series. Dutton (2000) authored a tutorial for the completion of the OSS procedure. Krapohl (2002) provided an update to the OSS normative data, using hand-scoring practices used by examiners trained at the Department of Defense Polygraph Institute; the OSS method remained unchanged at that time.

Krapohl and McManus (1999) reported they satisfied all of their development objectives with the exception of expediency, in that the OSS required some time investment to obtain the physiological measurements. While the earlier OSS required more time to complete compared with traditional pattern recognition approaches to field polygraph hand-scoring, the value of a reliable, well documented, measurement-based, and non-proprietary scoring procedure was not lost, and polygraph instrument developers recognized that deficits in expediency were easily remedied through software automation. The result has been that OSS became a computerized scoring algorithm. Presently three of four manufacturers of computer polygraphs sold in the US have included OSS in their software packages.

Despite its demonstrated efficiency with single issue ZCT polygraph examinations, the practical utility of OSS versions 1 and 2 was limited by the cumulative data structure, and by decision policies that do not attend to the complexities of multi-facet and mixed-issues examinations. The distributions of truthful and deceptive total scores from previous OSS versions are contingent on the number of question presentations regarding a single issue of concern, and on the number of test charts. Total scores are vulnerable to missing or uninterpretable data, as well as to additional data. Therefore, decision norms for earlier OSS versions do not theoretically generalize well to examination techniques involving two or four target stimuli, and are unable to take advantages of the completion of three to five test series as described by Kircher and Raskin (1988), Senter and Dollins (2004), and Senter, Dollins, and Krapohl (2004).

A further limitation of the earlier OSS version is that its data model and decision norms cannot be applied to multi-facet examinations regarding a single known allegation in which the examinee may be truthful to some but not all investigation targets, or mixed-issue screening examinations regarding multiple investigation targets involving unknown incidents. These conditions represent a substantial portion of field polygraph activity, and constitute a need for decision models that can evaluate individual spot scores in addition to total scores.

Krapohl and Norris (2000) evaluated OSS with confirmed criminal investigation exams using the Modified General Question Technique (MGQT, Ansley, 1998; Weaver & Garwood, 1985), and observed that human scorers provided better sensitivity to deception than attempts to apply the total score decision model of the earlier OSS version to spot scoring conditions. Krapohl and Norris also observed that the OSS model outperformed human scorers in terms of specificity to truthfulness. The results of Krapohl and Norris are consistent with mathematical expectations pertaining to the application of a cumulative data model to spot scoring circumstances, in which in the distribution of spot totals, upon which deceptive conclusions are based, can be expected to differ substantially from the distribution of cumulative totals, upon which the OSS-3 method was normed.

Method

Polygraph Component Sensors

Component sensors include upper and lower pneumograph sensors, cardio sensor cuff, and electrodermal sensors. Pulse-oximiter components have been available for some time (Kircher & Raskin, 1988), though they are used less commonly and are not included in presently available computer scoring algorithms. Peripheral activity sensors have become required components in the context of increasingly available strategies intended to defeat the polygraph test. At present, peripheral activity data is not a scored component in hand-scoring or computer algorithms, but is used to confirm the presence or absence of somatic peripheral nervous system activity among the autonomic nervous system data.

Algorithmic Approach

Nelson, Handler, and Krapohl (2007) introduced a major revision to OSS (Krapohl & McManus 1999; Krapohl, 2002), which is now called the Objective Scoring System, version 3 (OSS-3). The data model for OSS-3 is based on the aggregation of data through standardized scores and weighted averaging instead of simple cumulation. The use of standard *z*-scores allows OSS-3 normative data to approximate the distribution of total and spot scores regardless of the number of stimulus targets or test iterations. This important difference makes the OSS-3 method and OSS-3 normative data potentially more widely applicable to a variety of polygraph techniques and polygraph testing circumstances.

The new algorithm uses the mean comparison value as suggested by Elaad (1999), and is similar to previous OSS versions in its use of a two-distribution Gaussian model that was described by Barland (1985). This is in contrast to the single distribution bootstrap algorithm of Honts and Devitt (1992), and the single distribution permutation model of MacLaren and Krapohl (2003). The new algorithm differed substantially from previous versions in its use of standardized values, weighted averaging, and the use of Bootstrap resampling to train normative data for feature standardization, and the two distributions of truthful and deceptive decision norms (see Krapohl, Stern & Bronkema, 2002, for an introduction to probability and distribution models as these concepts apply to polygraph scoring.)

Whereas Krapohl and McManus (1999) managed the asymmetry of R/C ratios through a nonparametric transformation to a uniform distribution of septile bins, we transformed the dimensionless R/C ratios to their equivalent, though symmetrically distributed, natural logarithms. The natural logarithms of the distribution of asymmetrical R/C ratios will become a symmetrical normal distribution with a mean of zero. R/C ratios between zero and one will become an infinite number of logarithmic values between zero and minus-infinity, while values greater than one will become an infinite number of logarithm values between one and infinity. Because lognormal R/C ratios are normally distributed, we are justified in forgoing the granular nonparametric septile transformation of previous OSS versions in favor of parametric statistical procedures that offer greater potential statistical power.

An additional transformation was included at this point. Field polygraph examiners are trained to interpret negative numbers as indicative of greater differential reactivity to target stimuli than to comparison stimuli, and to interpret positive numerical values as indicative of greater differential reactivity to comparison stimuli than investigation targets. Natural logarithms of R/C ratios will be inverse to these expectations. We therefore inverted the sign values of all ratios, so that field examiners who wish to understand the operation of the algorithm can continue to interpret sign values in traditional ways. Sign values for pneumograph data were not inverted, so that human examiners can use a common paradigm for evaluating the data. Data are further transformed by standardizing all values for each component, using normative parameters that were obtained through bootstrap training (Efron, 1982; Mooney 1997; Mooney & Duval, 1993).

Bootstrap training. Bootstrapping is a computer-intensive method of obtaining empirical distribution estimates of parameters such as median and confidence ranges. Under ideal circumstances population parameters such as mean and deviation values would be achieved through testing every member of a population. Because that is often infeasible, test developers depend on samples of data that are intended to be representative of the population on which a test will be used. Variability will always be observed in a sample or population, and it is assumed that some degree of randomness will always be present. As a result, test developers are always concerned about the representativeness of a sampling distribution, and the biasing effect of even small departures from normality. The classical solution to problems of normality and representativeness is to construct numerous sampling distributions from which to calculate the sample parameters, and then use the distribution of sampling distributions as more robust population estimates than could be obtained from a single sample. Modern alternatives to the challenges of constructing numerous sampling distributions involve the use of computer intensive models to gain maximum value from each sampling distribution.

Bootstrapping involves the construction of an empirical bootstrap distribution of resampled sets, with replacement, from the sample data. Resampling is the equivalent of pulling a number at random from a hat after shaking, or randomizing, those numbers and then returning each number to the hat and then re-shaking or re-randomizing the numbers before selecting each subsequent number. This process is repeated continuously to create a resampled distribution of size equal to the sample from which each random selection is drawn. The process of constructing resampled distributions is then repeated numerous times to construct a bootstrap distribution of resampled distributions. With each random case selection, the probability of selecting a case from within the normal range is dictated by the law of large numbers and the central limit theorem, which tell us that if we completed this process a large number of times, our parameter estimates will regress towards the mean of the population represented by the sample.

Bootstrapping can be employed in nonparametric and empirical distribution models and does not depend on normally distributed data. Bootstrapping does assume that sample data are representative of the population, and bootstrapping will not correct for sampling problems. Bootstrap distributions are found to be normally distributed when the underlying sample or population data are normally distributed. Bootstrapping methods can therefore provide robust population estimates for use in parametric statistics, and can also be used to evaluate data for normality.

While it would take a crew of interns several weeks to complete the numerous resampling iterations necessary to achieve bootstrap estimates, modern computers can use brute-force to execute an exhaustive number of iterations with comparative ease. It is not uncommon for bootstrapping experiments to involve 1000 or 10,000, or even more resampled distributions. For each resampled distribution, the statistical parameters of interest are calculated for each resampled set, and a bootstrap distribution of those statistics is constructed by repeating this process many times. It is anticipated that some numbers might be randomly selected more than once within each resampled set, while others may not be selected in all. By using trimmed mean estimates, bootstrap resampling can reduce the influence of outlier or extreme values, against which mean and standard deviation statistics are non-robust or easily influenced.

We created 10,000 resampled sets of size equivalent to the OSS development sample of confirmed ZCT cases (N=292), and calculated population mean and standard deviation estimates for the lognormal R/C ratios for each polygraph component sensor. These values were then used to standardize each of the lognormal R/C ratios in the training sample. Table 1 shows the normative values for the natural logarithms of component ratios, which were derived from the first training bootstrap. Our normative standardization differs from those of Kircher and Raskin (1988; 2002), Raskin, Kircher, Honts, and Horowitz, (1988), who used ipsative standardization of component measurements between all charts to achieve the same goal of algebraically canceling out the physical units of measurement and achieving a consistent metric for evaluating data between the several test charts.

Table 1. Bootstrap mean and standard deviation scores for lognormal R/C ratios by component

	Mean	Standard Deviation
Pneumograph	-0.0385	0.1071
Electrodermal	-0.0179	0.1898
Cardiograph	0.0193	0.4987

Reduction of upper and lower pneumograph data. After standardizing each of the lognormal R/C ratios in the training sample, we then combined the standardized lognormal ratios, for each test stimulus, within each test series. We retained the method of combining upper and lower pneumograph values from previous OSS versions, in which values of opposite numerical sign are set to zero while keeping the signed value of greater magnitude when upper and lower pneumograph values are of similar numerical sign. This method is theoretically capable of retaining more data than the practice of arbitrarily discarding data from one of the pneumograph sensors (Harris & Olsen, 1994), and may be more robust against behaviorally adultered or uninterpretable pneumograph data than averaging the two components.

Trimmed outliers. Before combining the lognormal component ratios within each test chart, we first trimmed all ratios determined to be outliers according to a 3.8906 ipsative standard deviation boundary per each component. This meant that most data values would be considered usable and interpretable, while data values beyond greater than 99.99 percent of other values would be regarded as outliers.

Weighted Averaging. Instead of aggregating the component values through the simple addition methods of previous OSS versions, we combined those values within each test series, for each presentation of each target stimulus through weighted averaging. Several studies have suggested the electrodermal component provides the greatest contribution to diagnostic accuracy (Capps & Ansley, 1992; Harris & Olsen,1994; Kircher & Raskin, 1988; Raskin, Kircher, Honts, & Horowitz, 1988). Kircher, Kristjansson, Gardner, & Webb (2005) showed that cardiograph data is marginally more strongly correlated with the criterion than pneumograph data. Krapohl and McManus (1999) found that weighting the electrodermal component more strongly than cardiograph and pneumograph could reduce inconclusive results without compromising decision accuracy. Earlier OSS versions used integer weighting in which the numerical values assigned to electrodermal data were multiplied by two, meaning that one-half of the total cumulative score from the three component sensors (pneumograph, electrodermal, and cardiovascular) came from the electrodermal channel. We retained the use of integer level weighting, but differed from previous OSS versions in that electrodermal values were multiplied by three, while cardiograph and pneumograph data were multiplied by two and one respectively. The effective result is that component contributions are weighted in the following proportions: electrodermal = .5, cardiograph = .33, and pneumograph = .17. After first aggregating data for each test stimulus within-chart, as just described, we then aggregated the data between the three test series, by averaging the weighted mean scores for each spot. Mean lognormal R/C spot ratios were then further averaged together for a grand mean of standardized lognormal R/C ratios.

Bootstrap decision norms. We completed a second bootstrap of 10,000 resampled sets of the transformed data from the training sample (N = 292), and obtained population mean and standard deviation parameter estimates for separate normative distributions of truthful and deceptive persons for use in a two-distribution Gaussian signal detection model (Wickens, 1991; 2002). Figure 1 displays the quantile-quantile plots which verify that the bootstrap mean and standard deviation estimates for the weighted mean of standardized lognormal ratios of deceptive and truthful subsets are sufficiently normally distributed to justify the use of a parametric z-test in our decision model. Table 2 lists the normative parameters for confirmed truthful and deceptive cases in the training sample.

The ideal way to compare an individual score to those from deceptive or truthful groups would be to have access to the scores of every single deceptive or truthful person. Because that is unrealistic and impractical, test methods are often designed around samples of representative persons from truthful and deceptive groups. If we know how the distribution of scores in those groups (i.e., distribution shape, mean, and variance), then we can use statistical estimates to determine group assignments.

Kircher and Raskin (1988; 2002), and Raskin, Kircher, Honts and Horowitz, (1988) described an empirical Bayesian scoring algorithm that uses maximum likelihood estimates to assign cases to the group which a score most likely belongs to. Another method is the Gaussian-Gaussian signal detection model (Wickens 1991; 2002) described by Barland (1985), in which a score is compared to alternate normative distributions through the use of a simple hypothesis test. OSS-3 is constructed around this method, and assigns a case to the alternate category when the probability is very low, regarding inclusion in one of the normative groups.

The effectiveness of this model depends, in part, on the representativeness of the normative data, an accurate understanding of the distribution shape, and the robustness of our population parameter estimates (i.e., mean, standard deviation). A preferred method of understanding population parameter estimates is to calculate the estimates from a distribution of the mean and variance estimates of numerous sampling distributions, thereby reducing the influence

Figure 1. Quantile-quantile plots for bootstrap mean and standard deviation of weighted mean of standardized lognormal ratios

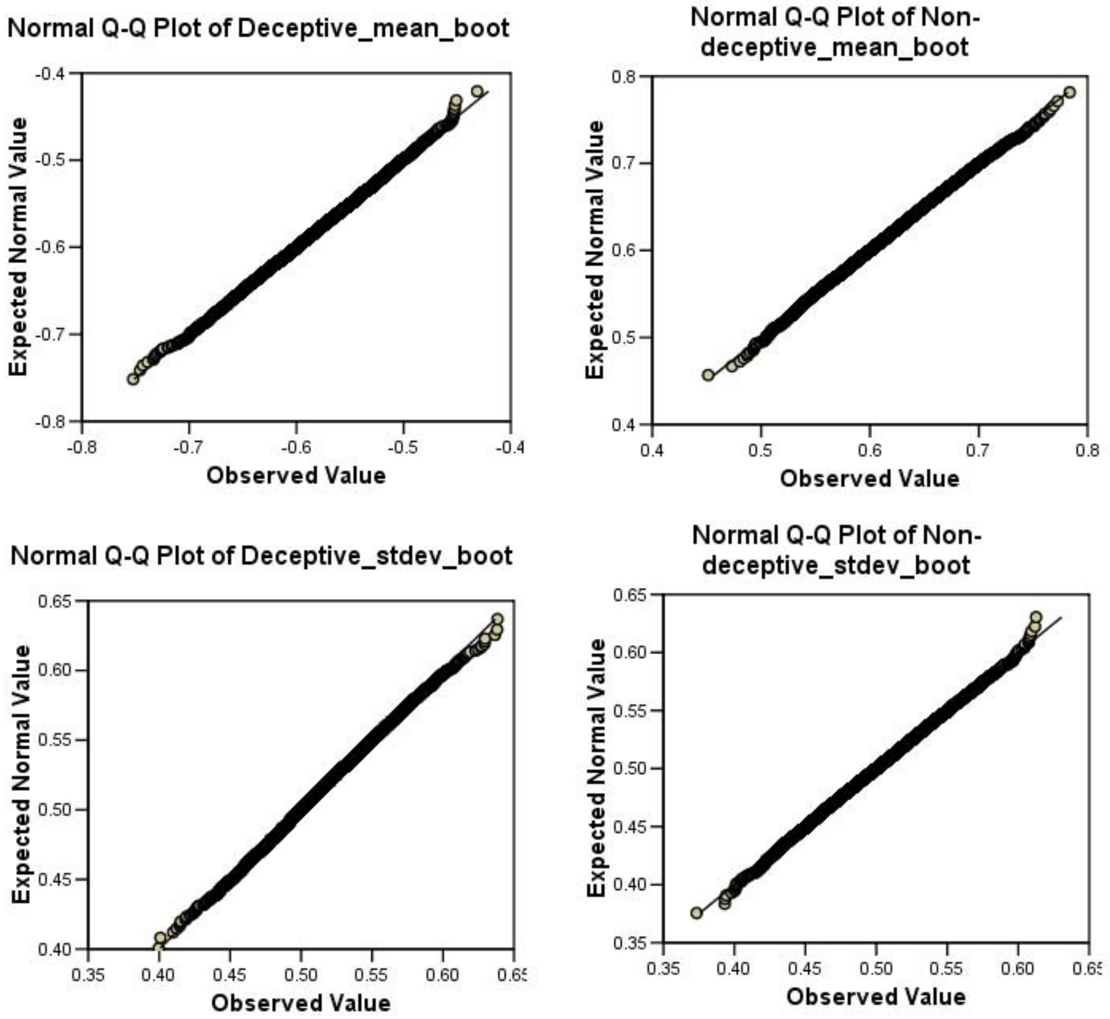


Table 2. Bootstrap mean and standard deviation scores for deceptive and truthful cases in the training sample.

	Mean	Standard Deviation
Deceptive	-0.5863	0.6192
Truthful	0.5188	0.5030

of bias from a single sample. This method depends on access to numerous samples of representative data. A modern alternative is to use brute-force computing and large scale bootstrap resampling methods to reduce the influence of bias in the calculation of population mean and variance estimates.

P-values. Using the bootstrap population norms for the grand mean of standardized lognormal R/C ratios, we evaluated the results of each examination against the distribution of confirmed deceptive cases, using a simple *z*-test that provides an estimated proportion of confirmed truthful persons that would produce a similar p-value. In field operation, cases are classified as truthful when the resulting p-value is less than the specified decision alpha. Whereas polygraph hand-scoring systems use point totals and cutscores to base decisions, statistical decision models based on signal detection theory make decisions according to alpha thresholds that are analogous to cutscores and represent a predetermined tolerance for risk of error. Very low p-values, when compared to the distribution of confirmed deceptive cases, would alert a field examiner that there is a low probability that the score was produced by a deceptive individual. It is therefore a matter of statistical inference that the individual was most likely truthful regarding the examination target.

Alpha decision cutpoints. Alpha thresholds are matters of administrative policy and tolerance for risk or error, just as much as they are matters of science. Common alpha thresholds are .05, .01, .001, and .1, which represent estimated decision error rates of 1 in 20, 1 in 100, 1 in 1000, and 1 in 10, respectively. Researchers in the social sciences commonly use .05 as a default or arbitrary boundary for statistical significance. Other alpha levels are employed as circumstances warrant. Because OSS-3 uses a two-distribution Gaussian signal detection model (Wickens, 1991; 2002), alpha thresholds for OSS-3 decision must be set for both truthful and deceptive classifications. Because the two alpha boundaries are set independently, they can be set asymmetrically, in order to optimize decision efficiency and balance sensitivity and specificity. Optimal alpha thresholds will served to maximize the correct classification of cases, while constraining inconclusive and erroneous results to acceptable levels. Using data from the training sample (N = 292) we determined that alpha = .1 presented an optimal condition for truthful classifications, including improved specificity to truthfulness and reduced inconclusives, while maintaining a minimal level of false-negative errors.

Two-stage decision policies. By default, OSS-3 uses two-stage decision policies (Senter, 2003) in which truthful classifications are attempted first, followed by attempts to classify cases as deceptive when they cannot be classified as truthful. If a case remains inconclusively resolved after that attempt, a second stage of decision policies is enacted, in which the between-chart mean of standardized lognormal ratios for each spot is assessed. Because the data are combined through averaging and standardization, the distribution and variance of each spot can be approximated by the distributions of grand mean values (see Table 2), which is unaffected by the number of test charts.

When an observed p-value is not less than the specified alpha, compared to the distribution of deceptive individuals, using alpha = .1, the grand mean of standardized lognormal values is then compared to the distribution of confirmed truthful cases, , using alpha = .05 and the same *z*-test procedure as before. When the resulting p-value is less than the specified decision alpha, it is interpreted as meaning there is a sufficiently low probability the score was produced by a truthful person and a case will be classified as deceptive.

Multiple comparisons and inflated alpha. In the case of single issue ZCT polygraphs, it is inconceivable that a subject could lie to one target stimulus while being truthful to others, or vice versa. Test stimuli are therefore non-independent, or dependent, and the addition rule allows us to calculate inflated alpha levels as $\alpha = \alpha_{\text{per test}}$ x number of tests. This means that while using alpha at .05 for single issue ZCT exams involving three non-independent target stimuli, the inflated alpha level is .05 x 3 = .15. In the case of multiple significance tests that are independent, (i.e., multi-facet or mixed-issues polygraph exams in which it is conceivable that a test subject could lie to one or more target issues while being truthful to one or more other investigation targets), the inflated alpha level can be estimated through the use of the multiplication rule, as $\alpha = 1 - (1\text{-}\alpha_{\text{per test}})^{\text{number of tests}}$. So, with a multi-facet or mixed-issues polygraph involving three independent targets, the inflated alpha is calculated as $1 - (1 - .05)^3 = .143$. Polygraph exams that use four questions will find the inflated alpha levels even higher. It is important to recognize that polygraph scoring schemes and decision policies based on integer point totals and cutscores are no less immune from multiple comparison and alpha complications. The effects of the addition rule and non-dependency will also play a role in the estimation of the likelihood of inconclusive test results in spot scoring circumstances.

In field polygraph testing, decision policies that neglect to correct for inflated alpha can be expected to contribute to decreased specificity to truthfulness and an increased false-positive error rate. The obvious benefits of completing multiple significance tests and using two-stage rules are the reduction of inconclusive results and improved sensitivity to deception. In the case of multi-facet and mixed issues examinations involving several independent investigation targets, there is also a semantic increase in sensitivity to a broader range of concerns.

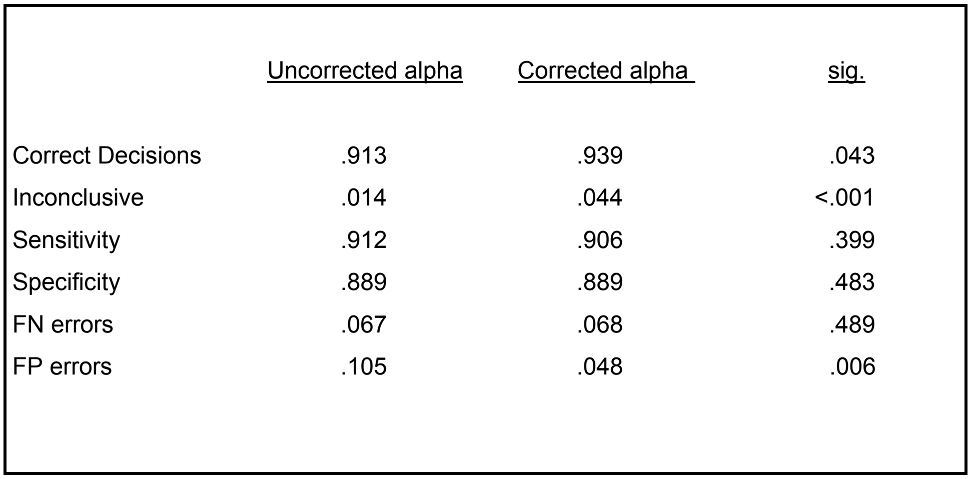
A number of statistical and mathematical procedures have been developed to correct for or reduce the impact of inflated alpha levels when completing multiple significance comparisons. The use of a Bonferonni correction to the specified alpha is one of the simplest methods, and applies to both dependent and independent circumstances. Bonferonni correction can be applied to a specified alpha level by multiplying the specified alpha by the number of comparisons. In polygraph spot scoring circumstances involving three stimulus targets, Bonferonni corrected as .05 x 3 = .0167. Senter (2003), and Senter and Dollins (2002) investigated spot scoring and total score decision policies and recommended the adoption of field practices that would serve to manage these known concerns through procedural solutions. It would, however, make equally good sense to begin to describe these concerns using the language of statistical inference that is common to other sciences.

Bonferonni correction. To avoid the increased likelihood of a type-1 error, in the form of false positive results, when completing the second stage of the two-stage scoring rules, we use a Bonferonni corrected alpha during the second stage of the two-stage decision policies. Inflation of the alpha is a known complication in any experimental or testing setting in which multiple simultaneous tests of significance are employed on the same data. With a single test of significance, using alpha at .05, there is a 5% chance (approximately 1 in 20 times) the data will result in a type 1 error and will appear significant due to chance alone. In practice circumstances, type-1 errors are called *false positives*. When conducting multiple simultaneous significance comparisons there is a mathematical inflation of the specified alpha. Calculation of the inflated alpha is typically done by one of two methods, depending on whether the various stimulus targets or investigation issues are independent, or non-independent/dependent.

We found that using two-stage rules (Senter, 2003) improved the sensitivity of the algorithm to deception from .828 to .913, with a corresponding reduction in inconclusive results, from 10.6% to 1.4%. Specificity to truthfulness remained constant at .89. However, the increase in sensitivity was not without cost, as the false positive error rate increased from 1.4% to 10.5%.

The application of a Bonferonni correction to the decision alpha reduced the false-positive rate to 6.5% with a minimal change in sensitivity to .906. Decision accuracy increased with the application of the Bonferonni correction, from 91.3% to 93.9%. To test the significance of these observed differences, we constructed a double-bootstrap Bonferonni t-test. Our double-bootstrap consisted of resampled sets of N = 292 cases from the training sample, from which we calculated mean estimates, before creating an secondary 292 resampled sets for each of the 292 resample sets in the primary bootstrap. The secondary bootstrap was used to calculate variance estimates which we used to complete a series of student's t-tests, using a Bonferonni corrected alpha, due to our use of multiple simultaneous significance tests. Table 3 shows the results of a double-bootstrap Bonferonni t-test. The increase in decision accuracy was significant at $p < .05$, but not significant when compared to the corrected alpha of .008. The reduction in false positive errors was significant, as was the increase in inconclusives. Despite the increase in inconclusive results, the overall inconclusive rate of 4.4% was regarded as tolerable in consideration of the increased decision accuracy and decreased false-positive error rate.

Table 3. OSS-3 (two-stage) results with and without Bonferonni correction.



	Uncorrected alpha	Corrected alpha	sig.
Correct Decisions	.913	.939	.043
Inconclusive	.014	.044	<.001
Sensitivity	.912	.906	.399
Specificity	.889	.889	.483
FN errors	.067	.068	.489
FP errors	.105	.048	.006

Experiment 1

Receiver operating characteristic. Using the receiver operating characteristic (ROC), we calculated the area under the curve (AUC) for OSS-3 and OSS-2. ROC statistics have the advantage of reducing several dimensions of accuracy concern, including sensitivity, specificity, and error rates, to a single numerical value. This makes it possible to easily compare the efficiency of different methods, both numerically and graphically. The advantages of this method become obvious when considering the ease of comparing two numbers compared to that of comparing separate tables of values. Because they evaluate decision accuracy across all possible decision cut-points, ROC statistics can provide analysts and decision makers with estimates of classification efficiency that are more easily integrated into decisions regarding tolerance for risk, compared with the challenges of generalizing accuracy estimates based on tables of values for varying cut-points, or the limitations of a single arbitrarily established cut-point. ROC estimates offer an important advantage over Bayesian estimates in that they are more resistant to base-rate influence. ROC estimates can be thought of as the likelihood that a randomly selected case will be correctly categorized, using a randomly selected decision cut-point. Areas under the curves were AUC = .964 for OSS-3 and AUC = .971 for OSS-2 using the OSS training sample. Data are shown in Figure 2. Table 4 shows that the 95% confidence interval of .945 to .983 for OSS-3 does not differ significantly from the .956 to .987 confidence range observed for OSS-2.

Bonferonni test. We conducted a Bonferonni t-test, using a double-bootstrap distribution of the training sample (N = 292). OSS-3 provided overall performance that equaled or exceed that of OSS-2. The double-bootstrap consisted of 292 samples of the 292 cases in the training sample, for each sample of which we selected an additional 292 samples. Improvements were observed in sensitivity to deception, ($p < .001$), reduced inconclusive results ($p < .001$), and specificity to truthfulness, ($p = .048$) were significant ($p < .05$). Differences in sensitivity to deception ($p < .001$) and reduced inconclusives ($p < .001$) were significant using a Bonferonni corrected

Figure 2. Area under the curve for OSS-3 (two-stage) and OSS-2 (total score).

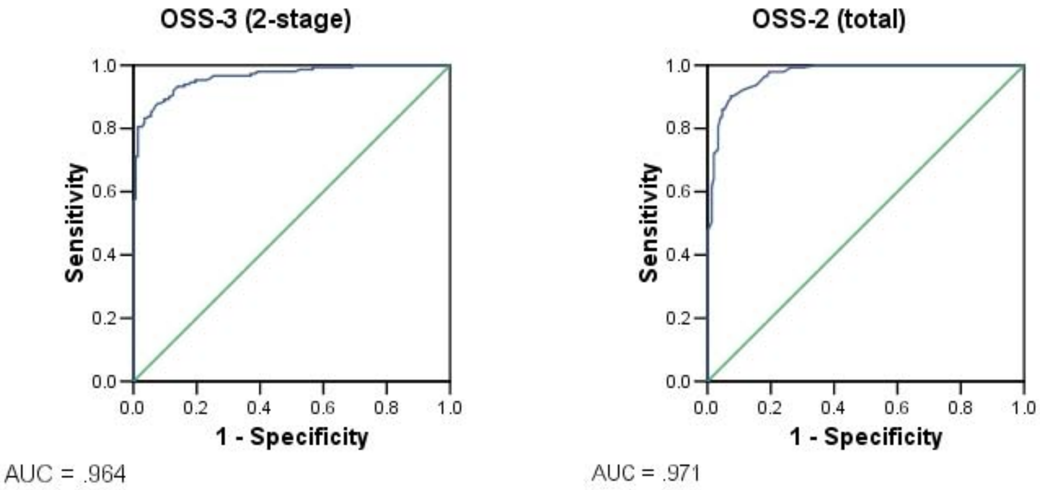


Table 4. AUC for OSS-3 (2-stage rules) and OSS-2 (total score) with training sample (N=292).

	Area	Std. Err.	95% Confidence Interval	
			Lower Bound	Upper Bound
OSS-3 (2-stage rules)	.964	.009	.945	.983
OSS-2 (total)	.971	.008	.956	.987

Table 5. Comparison of performance for OSS-3 and OSS-2 with training sample (N=292).

	OSS-3	OSS-2	sig.
Correct Decisions	93.9%	95.3%	.163
INC	4.4%	12.9%	<.001*
Sensitivity	90.6%	81.9%	<.001*
Specificity	88.8%	84.0%	.048
FN	6.6%	4.7%	.139
FP	4.8%	3.5%	.204

* denotes statistically significant improvement of OSS-3 over OSS-2.

alpha of .008, while the improvement in specificity was not significant at this level. Table 5 shows the results of the bootstrapped Bonferonni test.

Experiment 2

To further evaluate the new algorithm, we replicated the transformations and decision model of the empirical Bayesian algorithm, as described in (Kircher & Raskin, 1988; Kircher & Raskin, 2002; Raskin et al., 1988). That model is based on discriminate analysis and maximum likelihood estimation. Transformations of the empirical Bayesian method algorithm differ slightly from those of OSS-3, in that the empirical Bayesian method uses a *z*-score transformation to achieve a dimensionless measurement of differential reactivity to the test stimulus, whereas the OSS family of algorithms uses an R/C ratio to achieve the same objective (Krapohl, 1999). Transformations differ further in that the empirical Bayesian method uses ipsative standardization for each component, between all charts, and then averages data between charts for each component, before the calculation of maximum likelihood estimates that are then used to make posterior probability adjustments through a Bayesian probability model.

Because it is based on linear discriminate analysis, the empirical Bayesian method combines the between chart means of the component *z*-scores through addition, after weighting those *z*-scores with the unstandardized discriminate coefficients obtained from a discriminate analysis with the training sample (N=292). OSS-3 uses a normative standardization of lognormalized component ratios, and first aggregates data within each chart through weighted averaging of the component scores. Whereas the empirical Bayesian transformations produce a set of mean *z*-scores for all presentations of each test stimulus which are then further averaged for a grand mean *z*-score, OSS-3 transformations produce a weighted mean standardized measurement of differential reactivity for each presentation of each test. OSS-3 transformations then use unweighted averaging to combine the several presentations of each test stimulus for a set mean standard target scores, which are then further averaged for a grand mean score of the standardized lognormal ratios. We used SPSS (version 12.0) to calculate the discriminate function used in our replication of the empirical Bayesian decision algorithm. Table 6 shows the results the unstandardized discriminate coefficients and proportional component weight used in our replication of the empirical Bayesian algorithm.

Table 6. Unstandardized discriminate coefficients and proportional weights.

	Unstandardized discriminate coefficients	Proportional weight
Pneumograph	.629	.192
Electrodermal	1.735	.582
Cardiograph	.920	.280

Table 7 shows that the empirical Bayesian algorithm returned a decision accuracy rate of 94.4% with 7.5% inconclusive results. Sensitivity to deception was .879, while specificity to truthfulness was .865. False negative and false positive error rates were 4.0% and 6.3% respectively.

Table 7. Empirical Bayesian algorithm results with OSS training sample (N=292).

	Probability Analysis
Correct Decisions	94.4%
INC	7.5%
Sensitivity	87.9%
Specificity	86.7%
FN	4.0%
FP	6.3%

In the field of test development, results from a single sample or experiment cannot be regarded as adequately representative of how well a test method will work with the entire population. It is widely understood that accuracy estimates based on development samples are biased or optimistic estimates. Reasons for this include a variety of possibilities which include overfitting of the data model to the sample, reliability constraints with non-automated scoring, and the representativeness of the development sample. In general, simpler data models will not only tend to overfit less often, and will tend to provide greater interrater reliability among human scorers. For these reasons, accuracy estimates with validation samples are regarded as unbiased or less biased estimates.

Data were obtained from an archival sample that was constructed for a replication study conducted by Krapohl and Cushman (2006), which used earlier research (Krapohl, 2005) to develop Evidentiary Decision Rules for manual scoring of examinations conducted using the Zone Comparison Technique (Backster, 1963; Backster, 1990; Department of Defense Research Staff, 2006; Light, 1999).

Evidentiary decision rules (Krapohl, 2005; Krapohl & Cushman, 2006), are useful in field applications such as courtroom and paired-testing, or any testing context in which optimal balance of sensitivity and specificity and minimal inclusive results are among the highest priority. Krapohl and Cushman's sample consisted of N=100 event specific single-issue field polygraph exams, which were selected from an archive of confirmed cases, without regard for the original examiner's opinion. A more complete description of that sample can be found in previous publications. Those examinations were conducted using computerized polygraph systems. After extracting the Kircher features data (Kircher & Raskin, 1988; 2002; Dutton, 2000) using the Extract.exe software program (Harris, in Krapohl & McManus, 1999), we then scored of the replication sample (N=100) using the OSS-3 algorithm and the empirical Bayesian algorithm (Kircher & Raskin, 1988; 2002; Raskin et al., 1988) which were constructed using the open source spreadsheet application OpenOffice.org (available from Sun Microsystems), and a commercial spreadsheet from Microsoft. Artifacted and uninterpretable segments were not included in the computerized scores. Of the 1800 measurements, less than 2% of the data were marked as uninterpretable.

Table 8 shows the results of OSS-3 and the empirical Bayesian algorithm using the replication sample (N=100). Differences in decision accuracy was not significant (*p* = .365), though OSS-3 performed slightly better with 91% correct compared to 90.5% for the empirical Bayesian method. Difference in inconclusive results was significant (*p* = .002) using a Bonferonni corrected alpha of .008, with OSS-3 classifying 6.1% of the archival cases as inconclusive, compared to 15.0% for the empirical Bayesian method. OSS-3 returned fewer false negative errors than the empirical Bayesian method, with 8.1% compared to 12.2%, and more false positive errors, 7.9% compared to 4.0%, though those differences were not significant (*p* = .167) and (*p* = .117) respectively. OSS-3 showed greater sensitivity to deception, .858 compared to .778, which was not significant (*p* = .068). The empirical Bayesian algorithm showed fewer false positive errors, OSS-3 showed better specificity to truthfulness, .861 compared to .761, (*p* = .033) which was significant at .05, but not significant using a Bonferonni corrected alpha of .008. Figure 3 shows the Areas Under the Curve (AUC) for the Receiver Operating Characteristics for OSS-3 and empirical Bayesian algorithm to be .929 and .930 respectively.

To compare the results of the new algorithm to human examiners, we obtained the scored results from 10 human polygraph examiners, who scored the archival sample using evidentiary decision rules described by Krapohl and Cushman (2006). Evidentiary rules use two-stage decision rules (Senter, 2003), and employ cutscores that are empirically shown to reduce inconclusive results and improve specificity to truthfulness compared to traditional decision rules and cutscores.

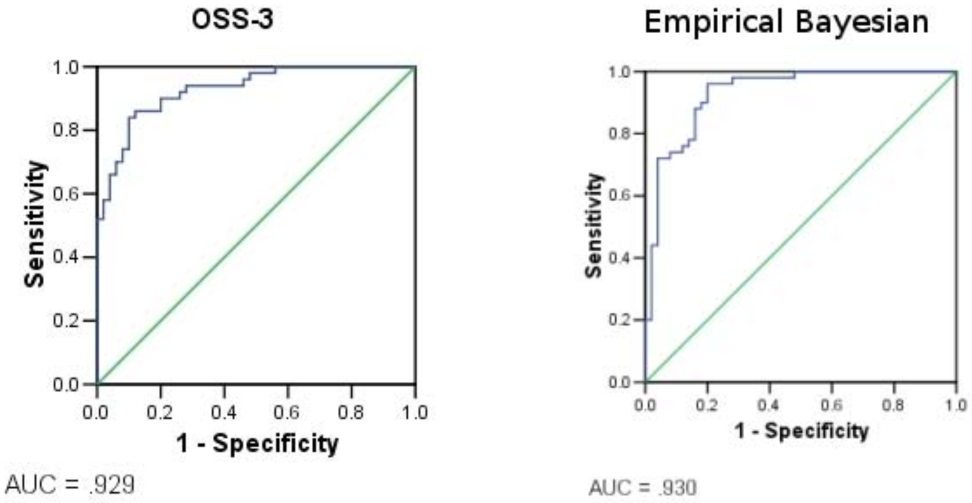
Participants. Human scorers were a self-selected cross-section of field examiners employed in private, law enforcement, and federal polygraph practice. Detailed information was not collected regarding the educational credentials and demographic background of the scorers. Human examiners ranged from 1 to 40 years in experience, with a median of 20 years (mean = 17 years). The examiners volunteered to score polygraph cases to verify their scoring abilities so to qualify for Marin protocol (paired testing) certification (Marin 2000; 2001). Human scorers were permitted to use a variety of existing scoring methods (cf., Backster, 1963; Backster, 1990; Bell, Raskin, Honts, & Kircher, 1999; Department of Defense Research Staff, 2006; Matte, 1996; Matte 1999; Handler, 2006).

Table 8. Comparison of performance for OSS-3 and the empirical Bayesian algorithm with the Marin replication sample (N=100).

	OSS-3	Empirical Bayesian	sig.
Correct Decisions	91.5%	90.5%	.365
INC	6.1%	15.0%	.002*
Sensitivity	85.8%	77.8%	.068
Specificity	86.1%	76.1%	.033
FN	8.1%	12.2%	.167
FP	7.9%	4.0%	.117

* denotes statistically significant difference using Bonferonni corrected alpha = .008.

Figure 3. ROC Area Under the Curve for OSS-3 and the Empirical Bayesian algorithm.



With the Krapohl and Cushman (2006) manual scoring data available we were afforded a benchmark against which to compare the performance of OSS-3. Because the Evidentiary Decision Rules (EDRs) for manual scoring led to the best overall accuracy, all comparisons here used those data with the understanding that the EDRs performance is probably higher than that found in common field practices. Table 9 shows decision accuracy and inconclusive rates for the 10 manual scorers and the OSS-3 algorithm. Decision accuracy rates for the human scorers ranged from 83.3% to 94.6% while inconclusive rates ranged from 4% to 13%. Ranked by decision accuracy, OSS-3 performed as well as or better than 9 of 10 human scorers.

Table 9. Rank order of 10 blind scorers and the OSS-3 algorithm by accuracy, in percent (N=100).

Rank	Scorer	Correct Excluding inconclusives	Total Inconclusive
1	2	94.6	7
2	**OSS-3**	**91.5**	**6**
3	1	89.9	1
4	4	89.7	13
5	9	87.4	5
6	6	86.7	10
7	8	86.7	10
8	5	86.5	11
9	3	83.5	9
10	10	83.5	3
11	7	83.3	4

Experiment 3

We then compared the maximum potential decision accuracy of OSS-3 to the 10 human scorers using ROC analysis. Figure 4 shows that the AUC = .878 for the average of the 10 human scorers in the Krapohl and Cushman (2006) sample. While OSS-3 outperformed the average of human scorers, inspection of the confidence intervals in Table 10 indicate that difference is not statistically significant.

Figure 4. ROC plots for OSS-3 and average of 10 human scorers with replication sample (N=100)

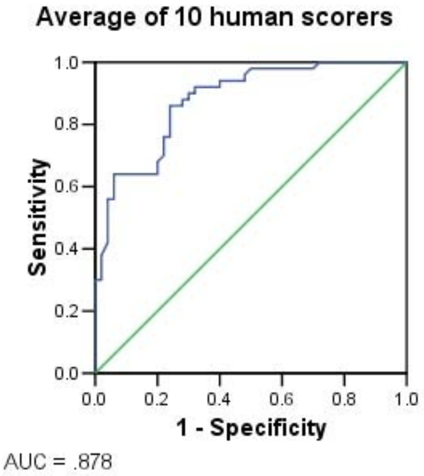


Table 10. AUC for OSS-3 (2-stage rules) and 10 human scores with Krapohl and Cushman (2006) replication sample (N=100).

	Area	Std. Err.	95% Confidence Interval	
			Lower Bound	Upper Bound
OSS-3 (2-stage rules)	.929	.024	.881	.976
10 human scorers	.878	.033	.813	.943

Monte Carlo methods. Next we used Monte Carlo simulation (Mooney 1997; Mooney & Duval, 1993) to compare the accuracy of the OSS-3 to the 10 human scorers. Monte Carlo methods are another class of brute-force computer-intensive methods of simulating the behavior of representative data, using massive sets of random numbers. We began by defining a Monte Carlo population space of N=1000 simulated examinations, for which we used random numbers to assign the confirmation status of each case according to an arbitrary base-rate of 0.5. Next, we use more random numbers to randomly assign the outcome of each deceptive or truthful case according to the proportions specified in Table 11, which are the averaged results of the 10 human scorers in the replication sample. Results for simulated deceptive cases were randomly assigned in the following proportions: correct = 0.792, error = 0.122, and inconclusive = 0.086. Results for simulated truthful cases were assigned according to the following proportions: correct: 0.824, error = 0.116, and inconclusive = 0.060. The use of random numbers for outcome assignment assured that the exact proportion of simulated results would never perfectly conform to the proportions observed in the replication sample, but would vary normally around the specified proportions with each iteration of the Monte Carlo simulation of the population space. Monte Carlo simulation techniques assume that randomness can never be completely eliminated, and uses normal variation in large-scale random simulations to observe how data can be expected to vary in live situations. We used 10,000 iterations of the Monte Carlo space of N=1000 simulated cases to calculate mean estimates and 95% confidence intervals for the 10 human scorers and used those values to calculate the significance of OSS-3 results compared to the human scorers.

OSS-3 produced an overall decision accuracy rate of 91.5% which was significantly better than the 87.2% average decision accuracy of the 10 human scorers (*z* = -3.91, *p* <.001). The 6.0% inconclusive results for OSS-3 was not significantly different from the 10 human scorers' average inconclusive rate of 7.2% (*z* = 1.56, *p* = .059). Table 12 shows overall decision accuracy and inconclusive rates and 95% confidence intervals from the Monte Carlo simulation, along with the OSS-3 computer algorithm results with the replication sample (N = 100).

Table 11. Averaged results, in percent, for 10 human scorers using Evidentiary Decision Rules.

	Deceptive Cases	Truthful Cases
Correct (with inconclusives)	79.2	82.4
Errors	12.2	11.6
Inconclusive	8.6	6.0
Correct (without inconclusives)	86.2	89.5

Table 12. Overall decision accuracy and (95% confidence interval) for all cases.

	Human Scorers	OSS-3	sig.
Correct (without inconclusives)	87.2 (85.0-89.3)	91.5 ***	<.001
Inconclusive	7.2 (5.7-8.9)	6.0	.059

*** p<.001

OSS-3 outperformed the human scorers with deceptive cases. OSS-3 showed a sensitivity rate of 86.0% which was significantly better than the average sensitivity level of 79.2% for the 10 human scorers (*z* = -3.75, *p* <.001), along with fewer false negative errors, 8.0% compared with the 12.2% (*z* = -2.86, *p* = .002), and fewer inconclusive results with 6.0%, compared with the 8.6% for the human scorers (*z* = -2.08, *p* = .019). Data are shown in Table 13.

Table 13. Decision accuracy and (95% confidence interval) for deceptive cases, in percent.

	Human Scorers	OSS-3	sig.
Sensitivity	79.2 (75.6-82.8)	86.0 ***	<.001
FN Error	12.2 (9.3-15.1)	8.0 **	.002
Inconclusive	8.6 (6.6-11.1)	6.0 *	.019

*** p <.001
** p <.01
* p <.05

The OSS-3 specificity rate of 86.0% was significant compared to the average of 82.4% (79.1% to 85.8%, *z* = 2.1, *p* = .018) for the 10 human scorers in the replication sample. OSS-3 also produced fewer false positive decision errors, with 8.0% compared with 11.6% (8.8% to 14.4%, *z* = -2.49, *p* = .006) for the averaged human scorers. Difference in inconclusive results was not significant for the truthful cases, with OSS-3 returning 6.0% inconclusive results, compared with 6.0% for the averaged human scorers (3.9% to 8.1%, *z* = .006, *p* = .502). Table 14 shows results for truthful cases.

Table 14. Decision accuracy and (95% confidence interval) for truthful cases, in percent.

	Human Scorers	OSS-3	sig.
Specificity	82.4 (79.1-85.8)	86.0 *	.018
FP Error	11.6 (8.8-14.4)	8.0 **	.006
Inconclusive	6.0 (3.9-8.1)	6.0	.502

** p <.01
* p <.05

Experiment 4

To further evaluate the effectiveness of the Kircher features and the possible advantages of a simplified scoring paradigm, we obtained the hand scored results from a cohort of seven inexperienced examiners in their eighth week of training at the Texas Department of Public Safety Polygraph School. These inexperienced examiners, all of whom have previous experience in Law Enforcement, had not yet completed their formal polygraph training and were provided a simplified rubric for polygraph scoring. The simplified hand scoring instructions employed only the three simple Kircher features, and instructions to score the cases in the Krapohl and Cushman (2008) replication sample (N = 100) using three-position scoring.

In an attempt to maximize interrater consistency, those instructions involve one primary scoring rule: the bigger-is-better principle in which any perceptible difference in magnitude between reactions to relevant and comparison stimuli is regarded as a scorable indicator of differential reaction. Two additional scoring guidelines were included in the simplified instructions provided to the inexperienced scorers. First, they were requested to refrain from assigning positive or negative point scores to erratic, artifacted and inconsistent response data, and instead assign a zero value, leave the score as blank or mark the score as artifacted. Second, the inexperienced scorers were instructed to score only those reactions that are timely with the stimulus, avoiding the assignment of negative or positive scores to reactions that occur prior to the stimulus onset or well after the end of the stimulus or point of answer.

Instructions were to score the cases visually, without the aide of measurement devices, using visual discrimination and conservative judgment as the arbiter of ambiguity in the case data. No explicit instructions were provided regarding the length of the scoring window, except a general instruction to score only those reaction segments which they were willing to argue as indicative of a Kircher feature and caused by the stimulus, while not due to artifact feature at the time of the examination. Artifact features include movement distortion, deep breath and other respiratory irregularities, and substantial instability of physiological response data.

Because we were interested in comparing non-mechanical Kircher feature scores with the computer algorithm, special instructions were provided for the interpretation of pneumograph data. Harris, Horner, and McQuarrie (2000) and Kircher, Kristjansson, Gardner, and Webb (2005), reported that Respiration Line Length, (Timm, 1982) provides a reliable approximation to changes in respiratory pattern that were correlated with the criterion of deception or truthfulness. Simplified scoring instructions defined three respiratory patterns as scoreable: 1) increase in respiratory baseline, of three or more respiratory cycles, before return to the pre-stimulus baseline, 2) suppression of respiratory amplitude of three or more respiratory cycles, following the stimulus onset, before return to the pre-stimulus level; and 3) slowing of respiration rate of three or more respiratory cycles from a consistent pre-stimulus level. Instability, movement, deep breaths, holding apnea, and all other features were to be scored zero or marked as an artifacted response segment.

The cohort of inexperienced scorers was instructed to refrain from formulating an opinion or conclusion regarding the truthful or deceptive status of the cases in the replication sample. Instead, we evaluated the mean and variance of the distributions of scores and determined decision cutscores using alpha boundaries common to social science research. Simplified scoring procedures resulted in a mean score of 8.85 for confirmed truthful cases (SD = 7.46), and a mean of -9.63 for confirmed deceptive cases (SD = 8.47). Table 15 depicts those data.

Our earlier experiments informed us that an asymmetrical alpha scheme could reduce the occurrence of inconclusive results among truthful subjects, with little effect on decision errors. We therefore selected scores similar to that used in OSS-3. Deceptive classifications would be made according to an alpha level of α = .05, scored against the distribution of truthful scores, while truthful classifications would be made at α = .1. To avoid an inflation of alpha, and a resulting potential increase in false-positive errors due to multiple statistical comparisons when using spot scores, we used a Bonferonni corrected alpha of α = .0167 for decisions resulting from a single spot. Using data in Table 16, we selected cutscores of +2 (α <= .1) for truthful classifications, and -4 (α <= .5) for deceptive classifications. For deceptive classifications based on spot scores, we used a Bonferonni corrected alpha of α = .05/3 = .017, because the cases in the replication sample included three relevant questions.

Table 15. Mean and standard deviations for truthful and deceptive cases, using the simplified scoring instructions.

	Average	St. Dev.
Confirmed Truthful (N=50)	8.85	7.46
Confirmed Deceptive (N=50)	-9.63	8.47

Table 16. Mean and standard deviations for truthful and deceptive cases, using the simplified scoring instructions.

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

Table 17 shows the results with the replication sample (N = 100) using the cutscores representing α = .1 for truthful classifications, α = .05 for deceptive classifications, and a Bonferonni corrected alpha of .017 for deceptive classifications based on spot scores obtained using the simplified scoring rubric. These results, obtained from inexperienced scorers, using the Kircher features on which OSS-3 is built, appear to rival those of the experienced scorers reported in Table 11.

We then completed another bootstrap resample of 1000 sets of the replication sample (N = 100), using the data from the 10 experienced using traditional scoring rules (Light, 1999) and seven inexperienced scorers, using the simplified hand-scoring rubric. Table 18 shows there are no significant differences between the results of the experienced scorers, using traditional hand-scoring systems, and inexperienced scorers who used a bare-bones scoring rubric consisting of Kircher features and simple rules.

Table 17. Results obtained with a 3-position hand-scoring rubric (N = 100) using only Kircher features, simplified scoring rules and inexperienced scorers.

	Simplified Hand Scoring
Correct Decisions	87.9%
INC	10.3%
Sensitivity (with inconclusives)	77.4%
Specificity (with inconclusives)	80.3%
Truthful correct (without inconclusives)	85.8%
Deceptive correct (without inconclusives)	90.1%

Table 18. Comparison of experienced scorers (traditional rules) and inexperienced scorers (simplified rules) (N=100).

	Experienced Scorers	Simplified Scoring	sig.
Correct Decisions	86.5%	87.5%	.348
INC	9.6%	10.2%	.416
Sensitivity	80.7%	77.6%	.299
Specificity	75.7%	80.2%	.221
FN	9.5%	12.9%	.225
FP	15.0%	8.9%	.091

To compare differences in inter-scorer consistency between the experienced scorers, and student scorers using a simplified hand-scoring system, we calculated the confidence ranges for Fleiss' kappa statistic for interrater reliability, using a final brute-force computerized statistical analysis, in the form of a two-dimensional double-bootstrap for which both cases and scorers were selected randomly to construct 100 x 100 resampled sets of the replication cases (N = 100). Inter-scorer agreement for the inexperienced scorers using the simplified scoring system (k = .61) had a slight but not significantly better performance advantage (p = .19, ns) over those of the experienced scorers (k = .57) whose reliability coefficient was identical to that reported by Blackwell (1999). Those results are shown in Table 19.

Table 19. Interrater reliability estimates for experienced scorers and inexperienced scorers using a simplified scoring system.

	Fleiss' kappa	95% Confidence Interval
Inexperienced scorers	.61	(.52 - .69)
Experienced scorers	.57	(.50 - .65)

Discussion

These data suggest that OSS-3 is capable of meeting or exceeding the capability of previous OSS versions and many human scorers along several dimensions, including sensitivity to deception, specificity to truthfulness, reduced false-negative and false-positive results, and reduced inconclusive results for deceptive cases. The average of human scorers did not out perform OSS-3 scores on any dimension. Equally important is that the new algorithm is based on mathematical transformations that can be theoretically applied to a much wider variety of examination techniques, including examinations consisting of two to four relevant questions and three to five test charts. The algorithm was designed to accommodate and manage the practical and mathematical complications inherent in multi-facet field investigation polygraphs. Design specifications for OSS-3 include specialized decision policies intended to optimize sensitivity and specificity with mixed-issues screening exams used in law-enforcement pre-employment testing and post-conviction offender testing programs.

We do not recommend increasing the number of investigation targets beyond four relevant questions, though it would be theoretically feasible to do so. Our reasons to advocate constraining the number of acceptable targets are based on the inescapable mathematical compromises necessitated by the effects of common statistical principles for dependent and independent probability events, which advise us to anticipate shifts in error rates that result from the inflation of the specified decision alpha with multiple test questions, and the increase in complex outcomes, including increased inconclusives, resulting from the correction of alpha to levels that would no longer well serve the purposes of field investigation. In short, the addition of more than four independent relevant questions incurs unavoidable errors as well as compromises to the validity of the test results. Constraining the number of investigation targets to four allows a range of flexibility that suits the needs of field polygraph investigators while retaining the ability to manage alpha decision boundaries responsibly.

As always, generalization and external validity of new methods and new knowledge is in part a feature of the representativeness of the normative development and validation sample, and there are known limitations pertaining to the application of polygraph techniques to low-functioning and psychotic persons – both populations which are overrepresented among criminal investigation and forensic subjects. We therefore encourage caution in the use of all polygraph methodologies with all exceptional persons.

The most accurate measure of the effectiveness of any decision model is practical experience in field settings. We conclude that the OSS-3 algorithm is capable of helping to meet the needs of field examiners and researchers, though we caution that the present study was limited to the effectiveness of the algorithm with event-specific/single-issue field investigation cases. Additional research is needed with multi-facet investigative polygraph examinations regarding known allegations, and with mixed-issues screening exams involving multiple investigation targets in the absence of any known allegations.

Although the mathematical transformations and statistical demands of OSS-3 are recognizably more involved than those of previous OSS versions, the procedure can be performed by computers with perfect consistency. Human scorers will never provide reliability that exceeds that of an automated procedure. We have no expectations that field polygraph examiners would attempt to calculate OSS-3 results by hand, but have endeavored to provide a complete description of the OSS-3 method for those who wish to study it. While OSS-3 provides results in the form of recognizable probability values, developers of existing hand-scoring systems in present use have not published or specified any tabular or mathematical methods for the calculation of the level of significance for hand-scored results. Instead existing hand-scoring systems are based on cumulative point totals that pertain to unspecified probability distribution models. Cut-scores for polygraph hand-scoring systems have been investigated for their empirical performance, and may be suboptimal compared with decision thresholds derived through methods based on statistical models. At present, little can be determined regarding how most polygraph cutscores conform to common alpha thresholds in similar signal detection models.

Our ability to study the existing range of computerized scoring algorithms is limited by incomplete documentation for the existing methods, and by proprietary and patent interests that preclude independent investigators, and field examiners from studying and completely understanding those methods. Another difficulty has been the lack of access or difficult access to raw data. We recommend that all manufacturers of polygraph field equipment make data available in a non-binary format that can be easily accessed by researchers equipped with common computer spreadsheets and statistical software. Minimally, all polygraph equipment manufacturers should export the Kircher measurements to a format that is easily machine readable. Presently three companies (Lafayette, Limestone and Stoelting) save these data in an accessible way.

A limitation of all presently available computer based scoring algorithms is that the physiological measurement data cannot be assumed to be robust against artifacts and data of compromised interpretable quality. There is no theoretical rationale suggesting that Kircher features, upon which OSS-3 is built, would be robust against data of marginal or unusual interpretable quality. Similarly, there is no published evidence that any of the features employed by any presently available computer algorithms would robust with uninterpretable data, or can effectively identify data of uninterpretable quality. Present methods of identifying artifacts through extreme values should be regarded as a blunt approach to the problem. Artifacted and uninterpretable data is simply uninterpretable, reminding us of the old adage in computer information processing "garbage in, garbage out." The inclusion of a Test of Proportions in the design specifications for the OSS-3 algorithm does not replace the need for further study in the areas of automated artifact and countermeasure detection. We remind the reader that human examiners should not yet rely on any scoring algorithm without carefully reviewing all test data for interpretable quality.

Research into automated polygraph scoring algorithms began in earnest during the 1980s, and automated algorithms have been available to polygraph examiners since the 1990s. However, there has been a general reluctance among examiners to base polygraph decisions on them. Raskin, Kircher, Honts, and Horowitz (1988) reported that discriminate analysis outperformed blind scorers but did not outperform original examiners. Honts and Amato, (2002) reiterated this conclusion. Honts and Devitt (1992), found no significant differences between the performance of expert human examiners, as original scorers, and the results of two automated algorithms, using discriminate analysis and bootstrapping, and suggested that bootstrapping outperformed the other methods and offered other advantages. Honts and Devitt also noted that their expert examiners were not representative of average field examiners. Research comparing human scorers to other automated algorithms has been mixed. Blackwell (1994), found that early versions of the Polygraph Automated Scoring System (PASS) (Harris &

Olsen, 1994; Olsen, Harris, & Chiu, 1994) did not perform as well in laboratory mock-crime experiments, though accuracy of the computer algorithm appeared to significantly improve with subsequent versions (Blackwell, 1996; 1998). Concerns about the representativeness of the study's human scorers apply to all previous comparisons of computerized and human scorers. The present study is an exception, and includes both experienced and inexperienced human scorers.

Just as the use of brute-force or computer intensive statistical analysis can facilitate our human understanding of the meaning and relevance of obscure physiological signals, the use of automated computer scoring algorithms can foster improvements to human scoring methods and human skills. Presently available computer scoring algorithms may not be capable of considering important nuances in the data as well as human scorers, though Kircher et al. (2005) suggested that original examiners do not seem to benefit from extrapolygraphic information. Nevertheless, there exists a need for further study in the areas of data quality and artifact detection. Standard field practice has been to rely primarily or even exclusively on manual scoring of the polygraph data. Criswell (2007) reported that the American Association of Police Polygraphists has declared it unethical for an examiner to base an opinion solely on the results of a computer scoring algorithm.

Because the mathematical scoring of polygraph test data is concerned only with the identification of statistical significance, we favor the wider use of field practices in which test results are described in terms of *significant reactions* or *no significant reactions* for all types of examinations. Holden (2000) previously discussed the difference between test results and professional opinions. While the presence or absence of statistically significant test results is a matter of objective mathematics, it remains the examiner's responsibility to ensure that nothing other than deception or truthfulness on the part of the examinee would cause those data to appear significant or non-significant. The determination that significant reactions are indicative of deception therefore remains a matter of both professional skill and the accuracy of the psychophysiological constructs that explain why people do or do not respond to polygraph test stimuli (see Handler & Honts, 2008). We do not suggest that a test itself should begin to replace professional responsibility or judgment but rather proffer the concept that algorithmic verification is in fact a useful tool for the field examiner.

Future research should also address the unknown limitations of automated physiological measurement in the presence of artifacted or unusual data quality. Other research should describe the normative distribution of truthful and deceptive scores for various hand-scoring systems, thereby facilitating a more informed statistical comparison of the capabilities of computer-based automated systems against hand-scoring systems. Improved understanding of polygraph hand-scoring norms will assist a variety of scientific investigators to more easily understand and evaluate polygraph decision models. For the present, we recommend further consideration and investigation of the OSS-3 algorithm as a viable scoring system for field use and quality assurance.

In consideration of evidence that OSS-3 and other computer scoring algorithms are capable of outperforming blind human scorers, the results of computer scoring algorithm should be considered carefully in quality assurance activities, though it will remain important for human examiners to review the data for adequacy for automated scoring until algorithms become available to automate those tasks. The use of automated algorithms for quality assurance purposes is less tenable with algorithms or hand-scoring measurements that employ proprietary or idiosyncratic physiological features, as differences between algorithm and human results are far more difficult to understand and resolve. We further recommend further investigation into the merits and possibilities of a simplified hand-scoring system based on Kircher features, simplified scoring guidelines and an empirically justified and statistically based understanding of decision rules and decision cutscores. In consideration of the effectiveness of the three Kircher measurements in both computerized and automated polygraph scoring systems, the use of idiosyncratic features and measurements, for which humans cannot easily understand or for which evidence of feature effectiveness is not available, is not justified.

We do not advocate the surrender of professional judgment to a computer algorithm, or the surrender of professional authority to any test method. Instead we recommend that field examiners, program administrators, and policy makers remain aware that professional judgment and professional ethics are domains of human concern for which there are formidable ethical complications when considering the implications of assigning responsibility for judgment to an automated process. Just as polygraph testing cannot completely substitute for an adequate field investigation, computer algorithms cannot substitute for inadequately administered examinations that suffer from poorly selected examination targets, ineffective linguistic construction, or test data of inadequate interpretable quality. Human judgments and policy decisions may be informed and improved by the results of testing and automated procedures, but the accuracy and effectiveness of those policies and judgment will depend in part on the abilities of those professionals to access complete documentation and data from research. We cannot justify the use of any algorithm with inscrutable features, transformation, decision policies and decision models. As with any evaluation measure, ethical use of a test or automated process requires a reasonable understanding of its design, development goals, and operations, including its strengths and limitations.

References

Ansley, N. (1998). The validity of the modified general question test (MGQT). *Polygraph*, 27, 35-44.

ASTM International (2002). *Standard Practices for Interpretation of Psychophysiological Detection of Deception (Polygraph) Data* (E 2229-02). West Conshohocken, PA: ASTM International.

Backster, C. (1963). *Standardized polygraph notepack and technique guide: Backster zone comparison technique.* New York: Cleve Backster.

Backster, C. (1990). *Backster zone comparison technique: Chart analysis rules*, Paper presented at the 25th annual seminar of the American Polygraph Association, Louisville, KY.

Barland, G. H. (1985). A method of estimating the accuracy of individual control question polygraph tests. In *Anti-terrorism; forensic science; psychology in police investigations: Proceedings of IDENTA-'85* (142-147). Jerusalem, Israel: The International Congress on Techniques for Criminal Identification.

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

Blackwell, N.J. (1994). *An evaluation of the effectiveness of the Polygraph Automated Scoring System (PASS) in detecting deception in a mock crime analog study*. Department of Defense Polygraph Institute Report DoDPI94-R-0003. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A305755.

Blackwell, N.J. (1996). *PolyScore: A comparison of accuracy*. Department of Defense Polygraph Institute Report DoDPI95-R-0001. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A313520.

Blackwell, N.J. (1998). *PolyScore 3.3 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations*. Department of Defense Polygraph Institute Report DoDPI97-R-006. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A355504/PAA. Reprinted in *Polygraph*, 28, (2) 149-175.

Capps, M. H. & Ansley, N. (1992). Numerical scoring of polygraph charts: What examiners really do. *Polygraph*, 21, 264-320.

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

Capps, M. H. & Ansley, N. (1992c). Analysis of federal polygraph charts by spot and chart total. *Polygraph*, 21, 110-131.

Criswell, E. (2007). Ethics: who really needs 'em anyway. *Police Polygraph Digest*, 1, 6-8.

Dollins, A. B., Krapohl, D. J. & Dutton, D.W. (2000). Computer algorithm comparison. *Polygraph*, 29, 237-247.

Dutton, D. (2000). Guide for performing the objective scoring system. *Polygraph*, 29(2), 177-184.

Department of Defense Research Staff (2006). *Federal Psychophysiological Detection of Deception Examiner Handbook*. Defense Academy for Credibility Assessment (formerly the Department of Defense Polygraph Institute). Ft Jackson, SC. Retrieved 1-10-2007 from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Capital City Press: Montpelier, Vermont.

Elaad, E. (1999). The control question technique: A search for improved decision rules. *Polygraph*, 28, 65-73.

Gordon, N. J. (1999). The academy for scientific investigative training's horizontal scoring system and examiner's algorithm system for chart interpretation. *Polygraph*, 28, 56-64.

Gordon, N. J. & Cochetti, P.M. (1987). The horizontal scoring system. *Polygraph*, 16, 116-125.

Handler, M.D. (2006) The Utah PLC. *Polygraph*, 35(3), 139-148.

Handler, M.D. and Honts C.R. (2008) Psychophysiological mechanisms in deception detection: a theoretical overview. *Polygraph*, 36(4) 221-236.

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

Harris, J. C., Olsen, Dale, E. (1994). *Polygraph Automated Scoring System*. U.S. Patent Document. Patent Number: 5,327,899.

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

Holden, E. J. (2000). Pre- and post-conviction polygraph: Building blocks for the future procedures, principles and practices. *Polygraph* , 29, 69-116.

Honts, C. R. & Amato, S.L. (2002). Countermeasures. In Murray Kleiner (Ed.) *Handbook of Polygraph Testing*. (251-264). Academic Press. San Diego.

Honts, C. R. & Devitt, M.K. (1992, August 24). *Bootstrap decision making for polygraph examinations*. Department of Defense Polygraph Institute Report DoDPI92-R-0002. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A304662.

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), 1-16.

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

Kircher, J.C., & Raskin, D.C. (1999). *The Computerized Polygraph System* (Version 3.0) Manual. Salt Lake City, UT: Scientific Assessment Technologies.

Kircher, J. C. & Raskin, D.C. (2002). Computer methods for the psychophysiological detection of deception. In Murray Kleiner (Ed.) *Handbook of Polygraph Testing*. :Academic Press: San Diego.

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

Krapohl, D. J. (1999). Proposed method for scoring electrodermal responses. *Polygraph*, 28, 82-84.

Krapohl, D. (2002). Short Report: An update for the Objective Scoring System. *Polygraph*, 31, 298-302.

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

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

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

Krapohl, D. J. & Norris, W.F. (2000). An exploratory study of traditional and objective scoring systems with MGQT field cases. *Polygraph*, 29, 185-194.

Krapohl, D. J. & Dutton, D.W. (2001). Respiration line length. *Polygraph*, 30, 56-59.

Krapohl, D. J., Dutton, D. W. & Ryan, A.H. (2001). The rank order scoring system: Replication and extension with field data. *Polygraph*, 30, 172-181.

Krapohl, D. J. & Stern, Brett, A. (2003). Principles of multiple-issue polygraph screening: A model for applicant, post-conviction offender, and counterintelligence testing. *Polygraph*, 32, 201-210.

Krapohl, D. J., Stern, B. A. & Bronkema, Y. (2002). Numerical analysis and wise decisions. *Polygraph*, 32(1), 2-14.

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

MacLaren, V. & Krapohl, D. (2003). Objective assessment of comparison question polygraphy. *Polygraph*, 32, 107-126.

Marin, J. (2000). He said/She said: Polygraph evidence in court. *Polygraph*, 29, 299-304.

Marin, J. (2001). The ASTM exclusionary standard and the APA 'litigation certificate' program. *Polygraph*, 30, 288-293.

Matte, J. A. (1996). *Forensic psychophysiology using the polygraph.* J.A.M. Publications: Williamsville, NY.

Matte, J. A. (1999). Numerical scoring systems in the triad of Matte polygraph techniques. *Polygraph*, 28, 46-55.

Miritello, K. (1999). Rank order analysis. *Polygraph*, 28, 74-76.

Mooney, C. Z. (1997). *Monte Carlo Simulation.* Sage Publications: Newbury Park, CA.

Mooney, C. Z. & Duval, R.D. (1993). *Bootstrapping. A nonparametric approach to statistical inference.* Sage Publications: Newbury Park, CA.

Nelson, R., Handler, M. & Krapohl, D. (2007, August). *Development and validation of the Objective Scoring System, version 3.* Poster presentation at the annual meeting of the American Polygraph Association, New Orleans, LA.

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

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.

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

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

Senter, S..M, & Dollins, A.B. (2004). Comparison of question series and decision rules: A replication. *Polygraph*, 33, 223-233.

Senter, S. M. & Dollins, A.B. (2002). *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.

Senter, S., Dollins, A. & Krapohl, D. (2004). A comparison of polygraph data evaluation conventions used at the University of Utah and the Department of Defense Polygraph Institute. *Polygraph*, 33, 214-222.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph*, 28(1), 10-27.

Timm, H. W. (1982). Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. *Journal of Applied Psychology*, 67, 391-400.

Van Herk, M. (1990). Numerical evaluation: Seven point scale +/-6 and possible alternatives: A discussion. *The Newsletter of the Canadian Association of Police Polygraphists*, 7, 28-47. Reprinted in *Polygraph*, 20(2), 70-79.

Weaver, R. S. & Garwood, M. (1985). Comparison of relevant/irrelevant and modified general question technique structures in a split counterintelligence-suitability phase polygraph examination. *Polygraph*, 14, 97-107.

Wickens, T. D. (1991). Maximum-likelihood estimation of a multivariate Gaussian rating model with excluded data. *Journal of Mathematical Psychology*, 36, 213-234.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York: Oxford.