Empirical Scoring System: A Cross-cultural Replication and Extension Study of Manual Scoring and Decision Policies

Mark Handler, Raymond Nelson, Walt Goodson and Matt Hicks

"The path of least resistance and least trouble is a mental rut already made. It requires troublesome work to undertake the alteration of old beliefs." — John Dewey

Abstract

A cohort of 19 international polygraph examiner trainees at the Texas Department of Public Safety Polygraph School used the Empirical Scoring System (Blalock, Cushman & Nelson, 2009; Krapohl, 2010; Nelson, Krapohl & Handler, 2008) to evaluate 100 confirmed event-specific criminal investigation polygraph examinations. Bootstrap analytic procedures were used to calculate accuracy profiles and statistical confidence intervals for test results comparing decision rules, including; the Grand Total Rule, Two-Stage Rules (Senter, 2003; Senter & Dollins, 2002 & 2004), Spot Scoring Rules, and traditional ZCT decision rules (Department of Defense Polygraph Institute, 2006). Bootstrap analysis of the distribution of trainee scores with the Empirical Scoring System resulted in a mean accuracy rate of 90.1% (95% CI = 83.8% to 95.8%), excluding 3.3% inconclusives (95% CI = 1.0% to 7.0%). A second bootstrap analysis of decision agreement showed that these inexperienced examiners demonstrated an average rate of agreement of 85% (95% CI = 65 – 97%). Evaluation of the distribution of sub-total scores revealed that 61% (95% CI = 51% to 70%), of the sub-total scores of truthful cases produced a non-positive score (a zero or negative value). Results from this study are consistent with those from previous studies (Blalock, Cushman & Nelson, 2009; Krapohl, 2010; Nelson, Krapohl & Handler, 2008), and provide further support for the validity of the principles inherent to the ESS, including the bigger-is-better rule, three position scoring, electrodermal weighting, two-stage decision rules, and the use of optimal cut-scores. The authors recommend continued interest in and additional research on the ESS as an expedient, valid and reliable method for manually scoring PDD examination data using statistical decision theory.

Acknowledgements

We are extremely grateful to Sabino Martinez, the Policia Federal in Mexico, Eva del Carmen Acua Tabares, Andres Aguilar Leon, Ivan Israel Albarran Santacruz, Miguel Angel Barcenas Monsivais, Lourdes Castro Garcia, Diego Armando Galindo Trujillo, Jose Manel Hurtado Garduño, Beatriz Irene Lopez Hernandez, Leslie Gabriela Montoya Arreola, Ana Silvia Von Schmeling Arechiga, Patricia Morales Luna, Jorge Zepeda Gonzalez, Claudia Anel Peña Alvarado, Juan Carranza Valencia, Zaida Ruth Alonzo Magaña, Miralva Garcia Ornelas, Omar Espino Lumbreras, Karina Barrios Arzate, and Isaac Ricardo Garcia Palacios. Without the commitment of these professionals none of this work would have been accomplished.

The authors thank Mr. Donald Krapohl, Mr. John Schwartz, Mr. Don Imbordino, Mr. Toby McSwain, Ms. Pam Shaw, Mr. Rick Kurtz, Mr. Geoffrey Flohr, Sgt. Robert Gilford, Mr. Chris Fausett, Mr. Mike Gougler, Dr. Stuart Senter, Dr. George Deitchman, and Dr. Charles Honts for their thoughtful reviews and comments to earlier drafts of this paper. The authors would grant unlimited use and duplication rights to any polygraph school accredited by the American Polygraph Association or the American Association of Police Polygraphists. Questions and comments are welcome at polygraphmark@gmail.com or raymond.nelson@gmail.com.

Introduction

The Empirical Scoring System (ESS) is a manual scoring model, first described by Nelson, Handler and Krapohl (2008). The developmental intent was to anchor every procedure and assumption used in the analysis of psycho-physiological detection of deception (PDD) examination data to empirical evidence and published scientific studies. A unique aspect of the ESS is that while it makes a strict demand for scientific proof and evidence for procedures and assumptions, the operational steps are quite simple compared to other manual scoring methods. The ESS employs a pattern-recognition approach using the on-screen data, and is completed visually, without the use of printing or any mechanical or automated measurements.

Psychologically, the ESS is based on the construct of salience (Handler & Nelson, 2007), which assumes that the magnitude of physiological responses to psychological stimuli are a function of the salience of those stimuli, and are mediated by a combination of emotional, cognitive, and behaviorally conditioned factors (Khan, Nelson & Handler, 2009). Salience, and the ESS make no assumptions about which exact emotion, articulate cognitions, or finite set of behavioral events do or do not serve as a basis for response to test stimuli. Instead all responses to stimuli are regarded as inclusive of some unknown proportion of each of these dimensional aspects of psychological response potential.

Physiologically, the ESS method of Test Data Analysis (TDA) requires analysis of only the primary reaction patterns derived from numerous studies on polygraph feature extraction (Dutton, 2000; Harris, Horner & McQuarrie, 2000; Honts & Driscoll, 1987, 1988; Kircher, Kristjansson, Gardner & Webb, 2005; Kircher & Raskin, 1988; Krapohl, 2002; Krapohl & McManus, 1999; MacLaren & Krapohl, 2003; Raskin, Kircher, Honts, & Horowitz, 1988; Nelson et al., 2008). Those reactions include phasic increases in skin conductance (or decreased resistance), increased relative blood pressure, and patterns of breathing movement often associated with movement suppression. Visual recognition of breathing movement suppression is accomplished through the evaluation of: 1) a suppression of waveform amplitude of three or more respiratory cycles following stimulus onset, or 2) a slowing of breathing rate for three or more respiratory cycles from a consistent pre-stimulus level, or 3) an increase in respiratory waveform baseline following stimulus onset and containing three or more breathing cycles before return to pre-stimulus baseline. This last pattern may or may not result from breathing movement suppression but has been shown to be a valid evaluation criterion (Kircher, Kristjansson, Gardner & Webb, 2005). Breathing apnea is regarded as the ultimate form of suppression (Department of Defense Polygraph Institute, 2006), but is easily faked and therefore scored only when it occurs at the relevant question.

The ESS does not employ rigid measurement periods or scoring windows, but requires that scores be assigned to reactions that are timely with, and caused by, the test stimuli. Early onset reactions are not scored nor are those with latencies that are atypically long for the examinee. Reactions that are obviously altered by movement, deep breath, or other voluntary or involuntary artifact event are also not scored.

The ESS makes no assumptions about, and places no requirements on, the linearity, scale or parametric shape of physiological response data. Instead, the ESS is based on the simple and robust assumption that larger reactions tend to occur in response to stimuli that are more salient due to dimensional factors that may include emotion, cognition, and/or behavioral conditioning. The ESS is based on the assumption that all observations or measurements of responses to test stimuli are estimates or approximations of the actual value of the response, and include elements of both systematic variance (i.e., data indicative of response to test stimuli) and uncontrolled variance (i.e., random measurement noise due to uncontrolled physiological, psychological, environmental or statistical measurement factors). The ESS further assumes that a more robust observation or measurement of responses to test stimuli can be achieved through the aggregation of multiple observations and measurements from several presentations of the test stimuli (e.g., *measure-twice cut-once* procedures in construction and wood-

working). Existing polygraph techniques have been developed around these assumptions, with several presentations of several versions of test stimulus questions which query the examinee's involvement in an allegation or issue of concern.

The ESS uses on-screen visual analysis to assign 3-position, nonparametric scores whenever a visibly perceptible difference in response magnitude is observed between pairs of relevant and comparison questions. Numerical scores are assigned for each component sensor, and a single composite score is assigned to the upper and lower pneumograph sensors as redundant measures of the same physiological response activity. The ESS does not make complex assumptions that the upper and lower pneumograph sensors somehow cancel, balance, or enhance each other. Instead, strength of reaction in pneumograph data is interpreted as a function of the frequency of occurrence of the scorable reaction patterns for the two pneumograph sensors. Several previous studies have shown electrodermal activity to be the most powerful and effective contributor to PDD examination results (Ansley & Krapohl, 2000; Capps & Ansley, 1992; Kircher & Raskin, 1988; Kircher et al., 2005; Krapohl & McManus, 1999; Nelson et al., 2008; Olsen, Harris & Chiu, 1994; Raskin et al., 1988). For this reason, all electro-dermal scores are doubled before calculating the sums for sub-total and total scores. ESS then uses simple addition to achieve weighted aggregate sub-total and grand-total scores for the several presentations of the test stimuli.

Categorical decisions of truthfulness or deception are made through statistical inference, using an equivariance-Gaussian decision model described by Barland (1985). This is accomplished by subjecting the sub-total and total scores to two-stage decision rules (Senter, 2003; Senter & Dollins, 2002; 2004). ESS scores are compared to cut-scores that are selected for a desired alpha boundary which represents a stated tolerance for error and required level of statistical significance and probability of error, based on normative data from Nelson et al. (2008). Decision alpha (cut-score) for deceptive results was set at .05, meaning that a test result would be considered statistically significant when the observed p-value (probability of error) is less than or equal to this level. Decision alpha (cut-score) for non-deceptive results was set at .1, meaning that a test result would be considered statistically significant when the observed p-value (probability of error) is less than or equal to this level.

Decisions based on sub-total scores present the well-known problem of inflated alpha, and corresponding increase in the potential for false-positive (FP) or type 1 errors when basing categorical decisions on multiple statistical comparisons regarding a single allegation or incident. Therefore, a Bonferroni corrected alpha of .017 (desired alpha of .05, divided by the 3 relevant questions) was used to reduce FP errors when basing decisions on any of the sub-total scores from the 3-question zone comparison tests (ZCT). If the number of sub-totals is different, then the correction factor is adjusted accordingly (for example, the correction factor for a 2-question ZCT would be .025). Because most inconclusive results tend to be truthful (Honts & Schweinle, 2009), the alpha for truthful decision was set at .1 in an attempt to balance sensitivity and specificity. Those interested in a more restrictive alpha for truthful case resolution may review the table in Analysis 4 to appreciate the changes in the accuracy profile when the alpha is adjusted via the cut-scores from .1 to .05 for truthful cases.

There are several advantages to selecting cut-scores based on normative data and statistical p-values, including the ability to select a cut-score that provides a desired level of decision accuracy or specified tolerance for error or risk, which is lacking in most current PDD hand scoring models in use today. Another important advantage of a decision theoretic and statistical approach to the selection of the PDD cut-score is that calculations of sensitivity and specificity levels, and their corresponding error rates, will be robust against difference in base-rates, such as in field settings where it is probably impossible to calculate the actual prior probability of deception; for example, a criminal suspect or examinee subject to polygraph screening. Knowing this ahead of time gives utility to the test result. In contrast, statistical metrics based only on Bayesian statistics or simple frequency calculations from sample data will be inherently non-resistant to differences in

base-rates and subject to legitimate criticism that they have poor generalizability to field situations in which the base rate of deception is unknown or expected to differ substantially from the circumstances of the research study.

Previous studies on the ESS (Blalock et al, 2009; Nelson et al, 2008) showed that inexperienced examiners, using a simplified empirically based manual scoring system of TDA were able to perform blind scoring tasks with decision accuracy, inconclusive rates, and interrater reliability that were statistically equivalent to those of experienced scorers (Krapohl & Cushman, 2006) using the prevailing and more complex 7-position TDA methods.

The present study is a cross-cultural replication of the original ESS experiment (Nelson et al, 2008), with a cohort of inexperienced polygraph examiner trainees from Mexico, who participated in training in the United States. In this study we:

1. Compared the accuracy profiles achieved by these international trainees to those achieved in previous studies (Blalock et al, 2009; Krapohl, 2010 ; Nelson et al, 2008);

2. Explored the level of interrater agreement among the participants in this study;

3. Investigated the use of two-stage decision rules (Senter, 2003; Senter & Dollins, 2002 & 2004) as compared to traditional, grand-total and sub-total (aka "spot score") rules;

4. Looked at the trade-offs of symmetric versus asymmetric alpha decision thresholds for truthful and deceptive cut-scores; and

5. Evaluated the prevalence of non-positive sub-total (spot) scores among truthful cases.

Method

Participants

A cohort of 19 police polygraph trainees in their eighth week of polygraph training at the Texas Department of Public Safety Law Enforcement Polygraph School, an American Polygraph Association accredited school, participated in the study. Participants were employees of the Policía Federal in Mexico, and were training to deploy in field environments in which PDD exams are used in the context of criminal investigations, and for the purpose of integrity and background screening of municipal police officers and police applicants. The ten female and nine male trainees all possessed a four-year college degree, equivalent with undergraduate education in the United States, in subjects including law, psychology, criminology and forensics. All were native Spanish speaking, and instruction was provided in Spanish by bilingual instructors from the United States.

Data Collection

Participants were instructed, in Spanish, in the use of the ESS, and then requested to evaluate an archival matched sample of 100 confirmed polygraph examinations that were randomly selected from the Department of Defense confirmed case archive. Fifty of the cases were conducted on examinees that were later confirmed as deceptive; the other 50 examinations were conducted on examinees that were later confirmed as non-deceptive to the investigative issue of concern. The same sample was previously used by Krapohl and Cushman (2006). All examinations were conducted using the Federal ZCT format (Department of Defense Research Staff, 2006), with three relevant questions and three test charts. Participants had received prior instruction in the current TDA procedures used by the National Center for Credibility Assessment (Department of Defense Polygraph Institute, 2006), and were asked to score the cases using the ESS, after approximately one hour of instruction on using the ESS model. Participants were asked to provide numerical scores only, and to refrain from making categorical decisions about the test results. Decision rules and cut-scores were established via normative data reported by Nelson et al (2008). Instructors who proctored the data collection phase were blind regarding the guilty status of each case.

Analysis 1 – Accuracy Profile

A Bootstrap Monte Carlo experiment was constructed to calculate the accuracy profile achieved by the study participants

using statistically optimal cut scores (alpha <0.1 for truthful and <.05 for deceptive, with Bonferroni corrected sub-totals) and using two-stage decision rules.

Results: Analysis 1 – Accuracy Profile.

Table 1 shows the mean and confidence intervals for the accuracy profile developed from a bootstrap resampling experiment of 1,000 iterations of the resample space of N = 100 sets of scores from the study participants. Bootstrap mean unweighted decision accuracy was 90.1% (95% CI = 83.8% to 95.8%), excluding 3.3% inconclusives (95% CI = 1.0% to 7.0%).

Table 1. Bootstrap Mean and Confidence Intervals for the ESS Accuracy Profile

	Result	95% Confidence Range
Proportion Correct	.901	(.838 to .958)
Inconclusives	.033	(.010 to .070)
Inconclusive Deceptive	.040	(.018 to .093)
Inconclusive Truthful	.039	(.017 to .091)
Sensitivity	.865	(.762 to .955)
Specificity	.881	(.782 to .961)
False Negative Errors	.103	(.024 to .192)
False Positive Errors	.089	(.021 to .174)
Positive Predictive Value	.906	(.818 to .978)
Negative Predictive Value	.895	(.800 to .976)
Unweighted Mean Accuracy	.901	(.839 to .954)

Discussion: Analysis 1 – Accuracy Profile

Test accuracy is a complex phenomenon composed of the interaction of several factors including among other things; construct validity, decision threshold and incidence rate. For this reason, it is not realistic to expect a single numerical index to adequately represent all of the dimensional variations that encompass the accuracy profile of a test or classification method. Instead, accuracy is most accurately understood through the evaluation of the various dimensions which determine the capability of a test to contribute incremental validity to practical decision making. Evaluation of multiple dimensional characteristics of test accuracy will allow developers to adjust testing protocols to optimize their testing objectives, and allows testing professionals, program administrators, and test consumers to make more effective use of the capabilities and advantages of the results from the PDD test.

Participants in this study produced results that were statistically equivalent to those achieved by previous studies on the ESS. Sensitivity and specificity rates were relatively balanced, as were inconclusive rates for truthful and deceptive examinations.

False positive and false negative errors were also found to be closely balanced in this experiment. Inconclusive rates observed during this experiment require additional explanation. Because of the randomization inherent to bootstrapping and Monte Carlo experiments, it is possible that some bootstrap or Monte Carlo distributions will result in zero inconclusives for some distribution. Inconclusive rates were calculated both within and between the truthful and deceptive groups. It is possible, under some randomized iterations, that there are zero inconclusives in one of the groups and not the other. When this occurs under exhaustive repetitions, the resulting between-group zero-inconclusive rate will be lower than the unweighted average of the within-group mean inconclusive rate, and will be more generalizable to field settings than the average of within-group inconclusive rates.

Analysis 2 – Interrater Reliability

We calculated the Fleiss Kappa statistic as a measurement of interrater agreement among the participants in the study. A two-dimensional double-bootstrap was calculated, for which both cases and scorers were selected randomly to construct 100 x 100 resampled sets of the participant scores (N = 100). Statistical confidence intervals were then constructed from the bootstrap distribution of scores.

To further illustrate the profile of interrater agreement achieved by the 19 study participants, we calculated the bootstrap distribution, including mean and 95% confidence range, for the proportion of agreement between decisions made by the study participants, using 1,000 iterations of the bootstrap resample space of 19 x 100 decisions.

Results: Analysis 2 – Interrater Reliability.

A moderate to substantial level of scoring agreement was achieved by the study participants, with k = 0.59 (95% CI = .52 to .65). However, the proportion of decision agreement observed among the participants was .84 (95% CI = .73 to .95). A bootstrap of the Pearson correlation coefficient among numerical scores was .84 (95% CI =.71 to .96), which was statistically significantly better than chance (p < .01).

Discussion: Analysis 2 – Interrater Reliability.

Interrater agreement among the inexperienced participants in this study was moderate to high, and was not statistically different from those observed in previous studies on the ESS (Blalock et al, 2009; Krapohl, 2010; Nelson et al, 2008). The ESS, using on-line evaluation, without mechanical measurements, outperformed previous reports of interrater agreement for experienced examiners (Blackwell, 1999) by a non-significant margin. These results were consistent with previous studies on the ESS (Krapohl, 2010; Nelson et al., 2008).

Analysis 3 – Decision Rules

To further investigate the influence of decision rules on ESS accuracy, additional analyses were conducted using 1,000 iterations of a bootstrap Monte Carlo model that was seeded with the scores from the study participants. Using statistically optimal thresholds (alpha < .1 for truthful and < .05 for deceptive, including Bonferroni correction to alpha for decisions based on sub-total scores), means and statistical confidence intervals were calculated for the accuracy profiles of ESS scores that were interpreted using different decision rules, including: the Grand Total Rule (GTR), Spot Scoring Rules (SSR), and traditional ZCT rules (TZR) (which involve the simultaneous use of the Grand Total and sub-total scores). Those results were then compared to ESS results using Two-Stage Rules (TSR).

Results: Analysis 3 – Decision Rules.

Table 2 shows the mean and confidence intervals for the different decision rules. The GTR produced the highest level of decision accuracy, however, differences in decision accuracy compared to the other rules was not significant. The GTR resulted in a significant increase in inconclusive results (p = .03) compared to the TSR. This difference loaded primarily on deceptive cases, but the overall change in inconclusives within the deceptive cases was not significant, nor was the corresponding reduction in test sensitivity to deception.

Table 2. Mean and confidence intervals for ESS2, TZR (Federal), GTR, & SSR

	ESS 2-stage	TZR	GTR	SSR
Proportion Correct	.901 {.837 to .958}	.870 (.17) {.789 to .942}	.914 (.36) {.853 to .968}	.875 (.22) {.792 to .946}
Inconclusives	.033 {.010 to .070}	.256 (<.01)** {.170 to .340}	.071 (.03)* {.030 to .130}	.285 (<.01)** {.200 to .370}
Inconclusive Deceptive	.040 {.018 to .093}	.091 (.05)* {.021 to .179}	.082 (.08) {.020 to .167}	.134 (<.01)** {.048 to .236}
Inconclusive Truthful	.039 {.017 to .091}	.426 (<.01)** {.259 to .625}	.065 (.16) {.018 to .140}	.441 (<.01)** {.275 to .628}
Sensitivity	.865 {.762 to .955}	.897 (.480) {.804 to .976}	.817 (.46) {.704 to .917}	.854 (.49) {.746 to .942}
Specificity	.881 {.782 to .961}	.398 (<.01)** {.262 to .536}	.880 (.48) {.783 to .961}	.397 (<.01)** {.260 to .539}
False Negative Errors	.103 {.024 to .192}	.027 (<.01)** {.017 to .063}	.103 (.49) {.023 to .196}	.027 (<.01)** {.017 to .060}
False Positive Errors	.089 {.021 to .174}	.181 (.03)* {.080 to .292}	.061 (.22) {.018 to .132}	.166 (.05)* {.067 to .260}
Positive Predictive Value	.906 {.818 to .978}	.832 (.05)* {.726 to .924}	.931 (.28) {.848 to .891}	.837 (.07) {.727 to .934}
Negative Predictive Value	.895 {.800 to .976}	.935 (.13) {.851 to .965}	.896 (.49) {.810 to .975}	.935 (.14) {.854 to .966}
Unweighted Mean Accuracy	.901 {.839 to .954}	.883 (.25) {.819 to .935}	.913 (.36) {.852 to .963}	.885 (.29) {.815 to .941}
*p < .05 **p < .01				

Compared to the ESS with TSR, the TZR, which include the simultaneous use of grand-total and spot-scoring rules, and require a positive score for each sub-total, resulted in statistically significant differences among several accuracy dimensions, including: increased inconclusives ($p < .01$) for both deceptive ($p = .05$) and truthful cases ($p < .01$), decreased specificity with truthful cases ($p < .01$) and increased false-positive errors ($p = .03$). Also, a statistically significant decrease was observed in positive-predictive-value ($p = .05$) when using the TZR. While most of the changes in accuracy resulting from the TZR were undesirable, one desirable change was observed in the form of a decrease in false-negative errors ($p < .01$).

The observed effect size for NPV for the (.935 - .895 = .040) was approaching, but did not achieve, statistical significance at the .05 level for the TZR. A post-hoc power analysis showed the power of the dimensional comparison to be ($\beta = .813$). A minimum statistical effect of .059 would be significant at the .05 level. A similar post-hoc power analysis on the percent correct achieved by the TSR and TZR indicated that the observed effect of .301 was achieved with ($\beta = .859$), while a minimum effect of .058, for decision accuracy, could achieve statistical significance at the .05 level. Post-hoc analysis of the effect size for unweighted accuracy (.018) revealed the power of the present experiment to be ($\beta = .907$), while a minimum effect size of .049 would be significant at the .05 level.

Spot Scoring Rules (SSR), using statistically optimal alpha cut-scores that were corrected for multiple within-test comparisons of deceptive and truthful scores, produced decreases in decision accuracy that were similar to the TZR and not significantly different from the other scoring conditions. The SSR resulted in statistically significant increases in inconclusives ($p < .01$) for both deceptive ($p < .01$) and truthful cases ($p < .01$), along with decreased specificity with truthful cases ($p < .01$) and increased false-positive errors ($p = .05$). The overall change in positive-predictive-value (PPV) was not significant ($p = .07$) at the .05 level, but was approaching statistical significance. A post-hoc power analysis indicates the power of the experimental dimension to be ($\beta = .690$). Like the TZR, the SSR did result in one desirable change, a decrease in false-negative errors ($p < .01$), likely a result of the requirement for all positive subtotals.

Discussion: Analysis 3 – Decision Rules.

Unweighted decision accuracy rates did not differ significantly among the four scoring conditions, and none of the scoring conditions produced a statistically significant difference in terms of test sensitivity to deception. The TSR produced a significant decrease in inconclusives compared to the GTR and TZR, along with significantly fewer FP errors and significantly greater PPV. The TZR produced significantly fewer FN errors than the TSR and GTR, at the cost of statistically significant increases in FP errors, and a very large significant effect for increase inconclusive results among truthful cases.

Different decision rules offer different advantages, constrain inconclusives, minimize certain types of error, or optimize specific dimensions of decision accuracy. Most of the differences in inconclusives appear to be due to the requirement for positive scores at all sub-totals in order to achieve a truthful result with the TZR. A significant increase in inconclusive results for deceptive cases for the TZR may be interpreted as a desirable change, because this dimensional change is related to the decrease in FN errors. A reduction in false negatives may be attainable with the TSR through the selection of a more conservative alpha decision threshold for truthful cases, though this is likely to result in an increase in inconclusives. This should be the focus of some future research.

Operationally, the difference between the TSR and TZR is that the TSR prioritizes the grand-total score first, regardless of the sub-total scores, and then only if inconclusive, proceeds to make deceptive classifications based sub-total scores. The TSR can be considered to emphasize balanced test sensitivity and test specificity, by making sequential use of the GTR and SSR, while the TZR prioritizes test sensitivity over test specificity, and amounts to the simultaneous use of the GTR and SSR. The TZR will permit a deceptive sub-total score to "trump" a truthful grand-total score, while the TSR regards the grand-total as more important than the sub-totals and will not allow a sub-total to "trump" the grand-total. The TZR

does not, however produce any increase in test sensitivity compared to the TSR, and the observed effect was limited to the reduction of FN errors at a cost of a loss of specificity and increased inconclusives. As always, practical decisions such as this are a matter of policies and operational priorities, just as much as they are a matter of science and decision theory.

These results show that the TZR is not more effective at catching liars than other decision rules. These results were obtained while using statistically optimal cutscores for all scoring conditions, so that any observed effect is not due to differences in decision cutscores and can be attributed to the decision rules. Readers should note that most, if not all, of the presently available and widely used scoring methods lack normative data and lack the ability to make inferential calculations of the probability of a test error. Field examiners, quality assurance reviewers, and program managers should be cautioned that using the ESS cutscores with other scoring methods is not recommended.

Based on these data, the TSR appears to be the optimal solution, with decreased inconclusives compared to the GTR. Use of the TZR should be restricted to circumstances that warrant a need for reduced false negatives, with a risk of a corresponding significant increase in inconclusives and a decrease in test specificity and positive predictive value and increased false-positive results. There appears to be no advantages to the use of the SSR with the ESS. Also, the SSR data reported here were calculated accounting for the deflation of alpha occurring with multiple within-test comparisons and optimal alpha cutscores. These precautions are not typically done in field settings and we predict uncorrected field-practice results will not improve the balance of test results. This too should be explored in future research.

Analysis 4 – Alpha Cut-scores

Using the ESS rules, we varied the decision alpha thresholds (cut-scores) to the effect of using a more conservative alpha for truthful cases. This should be of interest to those examiners who are concerned with risk-aversion and interested in a lower rate of false-negative (FN) results. As is common in many forms of testing, efforts to reduce errors may result in an increase in inconclusive results. We show the changes in the accuracy profiles for alpha held at < .05 for deceptive and varying alpha for the truthful from < .1 to < .05 in Table 3.

Results: Analysis 4 – Alpha Cut-scores.

FN error rates were reduced from the expected overall rate of ~.10 to ~.05 when we changed the alpha cut-score from .1 to .05. The difference in the rate of inconclusive results was significant ($p < .01$) and this difference was loaded on truthful cases, for which the difference was also significant ($p < .05$). Loss of test specificity within the truthful cases was statistically significant ($p < .01$) and Table 3 shows the results.

Discussion: Analysis 4 – Alpha Cut-scores.

This analysis compares decision thresholds in an effort to demonstrate the trade-offs encountered when a more stringent alpha is observed for the truthful cases, (equivalent to requiring a higher positive score to achieve a "No Significant Response" result). As can be seen, proportion correct, deceptive inconclusives, sensitivity, false positive and false negative results, positive and negative predictive value and unweighted accuracies do not differ significantly. However, imposing the more stringent threshold, results in increased inconclusive results for overall cases and especially for the truthful cases and a decrease in specificity. While the selection of alpha decision cut-scores is ultimately a matter of administrative policy as much as it is a matter of science, these results indicate that the current balanced approach of observing alpha at <.05 for deceptive cases and <.1 for truthful cases maintains a relatively high level of sensitivity and specificity, while holding the inconclusive rate low and constraining errors to tolerable proportions.

Analysis 5 – Proportion of Non-positive Sub-total Scores

To further evaluate the assumptions of the TZR, which require a positive sub-total score (spot scores) for all investigation target questions, bootstrap analytic procedures were used to calculate frequency, proportion and confidence intervals for the presence of non-positive sub-totals (i.e., sub-totals that are zero or negative scores) among the confirmed truthful cases.

Table 3. Results of varying the truthful decision threshold (alpha) with ESS rules

	ESS rules *truthful alpha < .1* deceptive alpha < .05	ESS rules *truthful alpha < .05* deceptive alpha < .05	Sig.
Proportion Correct	.901 (.837 to .958)	.904 (.839 to .957)	.488
Inconclusives	.033 (.010 to .070)	.095 (.040 to .160)	.005**
Inconclusive Deceptive	.040 (.018 to .093)	.072 (.019 to .149)	.141
Inconclusive Truthful	.039 (.017 to .091)	.121 (.038 to .229)	.016*
Sensitivity	.865 (.762 to .955)	.888 (.791 to .963)	.331
Specificity	.881 (.782 to .961)	.747 (.627 to .867)	.007**
False Negative Errors	.103 (.024 to .192)	.048 (.018 to .104)	.072
False Positive Errors	.089 (.021 to .174)	.131 (.041 to .234)	.166
Positive Predictive Value	.906 (.818 to .978)	.871 (.762 to .961)	.207
Negative Predictive Value	.895 (.800 to .976)	.940 (.870 to .978)	.134
Unweighted Mean Accuracy	.901 (.839 to .954)	.905 (.841 to .959)	.467
*p < .05 **p < .01			

Table 4. Frequency of non-positive sub-totals.

Questions	Proportion	95% CI
R5	13%	(71% to 91%)
R7	29%	(21% to 39%)
R10	37%	(27% to 46%)
Any RQ	61%	(51% to 70%)

Results: Analysis 5 – Proportion of Non-position Sub-total Scores.

Bootstrap analysis revealed that 61% (95% CI = 51% to 70%) of the truthful cases can be expected to result in at least one or more sub-total scores that are non-positive (i.e., zero [0] or negative scores).

Discussion: Analysis 5 – Proportion of Non-position Sub-total Scores.

Results of this experiment suggest that a large proportion of truthful persons will produce at least one non-positive sub-total score. This requirement results in a condition in which more than one half of truthful cases are regarded as incapable of being correctly classified, and the value of this rule (requirement for positive scores in all sub-totals) is therefore questionable. Some may assume this rule increases decision accuracy with deceptive cases, in terms of increased sensitivity to deception or decreased false negative errors. While it would be procedurally and mathematically impossible for this requirement to produce an increase in test sensitivity, this procedural requirement does result in a statistically significant reduction in false negative results, at the cost of a statistically significant increase in inconclusive results among truthful persons. A more practical, and precise, solution to the need for low false-negative error rates might be achieved through the selection of an alpha decision cut-score that assures the required level of precision with greater ability to constrain error rates. This should become the focus of a future study.

General Discussion

The trainees from the Mexico Federal Police demonstrated that ESS can produce balanced sensitivity and specificity, with no significant differences from results achieved during previous studies on the ESS (Blalock, Cushman & Nelson, 2009; Blalock, Nelson, Cushman, & Oelrich, 2010; Krapohl, 2010; Nelson et al., 2008). The inexperienced examiners (trainees) in this study scored polygraph charts at accuracy and reliability rates consistent with those of the experienced examiners reported by Krapohl and Cushman (2006) which should be of interest to trainers, field examiners and program managers. It seems reasonable to assume that field experience is valuable and contributes to increased skill and performance in test data analysis. Therefore, the performance of the inexperienced scorers might be attributable to an improved emphasis on empirically sound principles in their scoring method. Additionally, historical scoring exercises may have involved evaluators with considerable experience and expertise. Using these experts to test a scoring model is less likely to generalize to what will happen in the field. It is perhaps more informative to test the "weakest link in the chain" to estimate how well a model will work for the many, as opposed to the few. A final consideration is this system was taught to the students via a translator suggesting this simplified system is easy to communicate across language barriers.

Grand total decision rules were the simplest solution and provided the highest level of decision accuracy, though the difference was not significant. Two stage rules outperformed the grand total decision rules in terms of inconclusive results, and sensitivity to deception, with no significant difference in false-positive or false-negative errors. Traditional decision rules produced a significant increase in the rate of false - positive errors and inconclusive results for the truthful sample cases – a result that is consistent with previously published studies (Krapohl & Cushman, 2006).

Two-stage decision rules seem to provide more balanced test sensitivity and test specificity than traditional rules. While the traditional rules have served the profession well through the years, they may be sub-optimal. In addition it is becoming increasingly clear that traditional decision cutscores have not been studied in the context of normative data or correspondence with decision alpha levels. On the surface the traditional rules seem to benefit a "risk-aversive" testing program. Examiners with an inherent fear of a false-negative error may become convinced they would rather "interrogate and apologize" than allow themselves to be beaten. A closer consideration of this attitude reveals that in the end it may actually be detrimental.

Polygraph examiners and consumers of polygraph rely on the test's ability to differentiate the truthful from the deceptive. A

test with high sensitivity but poor specificity will have a low positive predictive value because of the high false-positive rate. The mathematical reality of this is that a lot of truthful subjects are classified as deceptive, and logic dictates, will be interrogated. Interrogating a truthful subject offers the opportunity for a number of negative outcomes, not the least of which could be a false confession. Also, field examiners traditionally pride themselves on their ability to secure a posttest admission, and examiners, their peers and their supervisors use this as a metric of success. Indeed, a number of organizations keep statistics on confession rates! Being unable to separate the truthful from the deceptive because of a test that is heavy on sensitivity and light on specificity is a set up for disappointment for the examiner, his or her supervisor, or consumer. While it may seem initially convenient to ensure test sensitivity at the cost of imbalance specificity, the long term effect will be corrosive of confidence among consumers. These consumers of polygraph rely on the diagnostic value of the test result to add incremental validity to the process in which the polygraph has been applied. Without diagnostic value, polygraph will be of no more value than computer voice stress analyzers.

The ESS model applies the principle of weighting the contribution of the EDA component most heavily, employs empirically supported two-stage decision policies, and uses statistically optimal thresholds (cut scores) that allow for error estimations and the dispensing of "lore-based" decision rules. The ESS is straightforward to use and easy to explain to polygraph examiners and non-examiners such as department administrators or adjudicators. ESS offers promising potential for gaining increased understanding and increased credibility among consumers of polygraph test results, with good consistently high criterion validity and interrater reliability that is as good or better than other scoring models.

A major advantage of the ESS, compared to current hand-scoring systems, is the existence of normative data that can be used to provide an understanding of the level of statistical significance achieved by various decision cut-scores (see Appendix A). In an era that emphasizes theoretically sound decision models, mathematically defensible results, and known methods for calculating the likelihood of an erroneous test result, all investigators involved in development and research of polygraph scoring systems should feel an obligation to publish normative data and significance tables for all polygraph scoring systems in present use.

No study is without limitations, and we note several limitations in this study. First, the number of evaluators contributing to this study is small and while, the sample itself is not small, it is also not large. This study addresses only Zone Comparison Technique polygraph results and does not attempt to address data collected using other polygraph techniques. We also realize that balance of sensitivity and specificity may not be the goal in all testing situations and there may be times when more strict or lenient tolerance exists for one type of error over another. This point is precisely why we advocate moving the profession towards results based on p-values, normative data, and the ability to compare a calculated probability of error to a stated tolerance for error. Polygraph professionals should strive to study and understand the normative data to the point where we may make reasonable estimates of our errors on individual cases. In this way we can more precisely predict the scientific strength of confidence in the results. We suggest others replicate this experiment to support or refute our findings in hopes that we can collectively improve the quality of polygraph for all.

As the late great social psychologist Leon Festinger (1987) stated in his remarks during the symposium Reflections on Cognitive Dissonance: 30 Years Later at the 95th Annual Convention of the American Psychology Association:

> No theory is going to be inviolate. Let me put it clearly. The only kind of theory that can be proposed and ever will be proposed that absolutely will remain inviolate for decades, certainly centuries, is a theory that is not testable. If a theory is at all testable, it will not remain unchanged. It has to change. All theories are wrong. One does not ask about theories, can I show that they are wrong or can I show that they are right, but rather

> one asks, how much of the empirical realm can it handle and how must it be modified and changed as it matures?

One thing is for sure: if we presently consider the polygraph to be either perfect or just as good as it need be, then there is no reason to study data or pursue improved methods of hand-scoring. If, however, we think of the polygraph as imperfect and capable of being improved upon, we must rise to the challenge of studying our theories and assumptions and be willing to release our grasp of any procedures which are arcane and suboptimal. Holding on for the sake of posterity, in the face of evidence and data that informs of ways to improve the accuracy profile of the polygraph examination will not only hold the profession back, it will be considered irresponsible and unethical by others with whom we share the social sciences. The authors recommend continued interest in, and additional research on, the ESS as an expedient, valid and reliable evidence based method for manually scoring PDD examination data using statistical decision theory.

References

Ansley, N. & Krapohl, D.J. (2000). The Frequency of appearance of evaluative criteria in field polygraph charts. *Polygraph*, 29 (2), 169-176.

Barland, G.H. (1985). A method for estimating the accuracy of individual control question tests. *Proceedings of Identa-85*, 142-147.

Blackwell, N.J. (1999). *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.

Blalock, B., Cushman, B. & Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. *Polygraph*, 38(4), 281-288.

Blalock, B., Nelson, R., Cushman, B. & Oelrich, M. (submitted 2010). Reliability of the Empirical Scoring System with Expert Examiners. A manuscript submitted to *Polygraph*, for consideration.

Capps, M. H. & Ansley, N. (1992). Analysis of private industry polygraph charts by spot and chart control. *Polygraph*, 21, 132-142.

Department of Defense Polygraph Institute (2006). Test Data Analysis: DoDPI numerical evaluation scoring system. Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007.

Department of Defense Research Staff (2006). Federal Psychophysiological Detection of Deception Examiner Handbook. available online: Retrieved 1-10-2007 from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf.

Dutton, D. (2000). Guide for performing the Objective Scoring System. *Polygraph*, 29, 177-184.

Festinger, L. (1987). Appendix B: Reflections on cognitive dissonance: 30 years later. In E. Harmon-Jones & J. Mills (Eds.), *Cognitive Dissonance-Progress on a Pivotal Theory in Social Psychology* (1999). Washington, DC: American Psychological Association.

Handler, M. & Nelson, R. (2007). Polygraph terms for the 21st century. *Polygraph*, 36, 157-164.

Harris, J., Horner, A. & McQuarrie, D. (2000). *An evaluation of the criteria taught by the Department of Defense Polygraph Institute for interpreting polygraph examinations*. Johns Hopkins University, Applied Physics Laboratory. SSD-POR-POR-00-7272

Honts, C. R. & Driscoll, L. N. (1987). An evaluation of the reliability and validity of rank order and standard numerical scoring of polygraph charts. *Polygraph*, 16, 241-257.

Honts, C. R. & Driscoll, L. N. (1988). A field validity study of rank order scoring system (ROSS) in multiple issue control question tests. *Polygraph*, 17, p. 1-15.

Honts, C. R., & Schweinle, W. (2009). Information gain of psychophysiological detection of deception in forensic and screening settings. Manuscript accepted for publication pending revision, *Applied Psychophysiology and Biofeedback*.

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., Kristjansson, S. D., Gardner, M. K. & Webb, A. (2005). *Human and computer decision-making in the psychophysiological detection of deception.* University of Utah. Final Report

Khan, J., Nelson, R., & Handler, M. (2009). An exploration of emotion and cognition during polygraph testing. *Polygraph*, 38 (3), 184-197.

Krapohl, D. (2002). The polygraph in personnel screening. In M. Kleiner (Ed.), *Handbook of Polygraph Testing* (2002). San Diego, CA: Academic Press.

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

Krapohl, D.J. (2010). Short Report: A test of the ESS with two-question field cases. *Polygraph*, 39(2), 124-126.

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

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

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

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

Raskin, D. C., Kircher, J. C., Honts, C. R. & Horowitz, S. W. (1988). *A study of the validity of polygraph examinations in criminal investigations*. Final Report, National Institute of Justice, Grant No. 85-IJ-CX-0040. 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. (2002). *New decision rule development: Exploration of a two-stage approach*. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC.

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

Appendix A

ESS – Monte Carlo Normative Data for Event-Specific ZCT Exams with 3 RQs
Mean deceptive score = -9.14 (SD = 8.74) Mean truthful score = 8.35 (SD = 7.89)

Truthful (NSR) Cut-scores	
Total NSR Cut-score	p-value (alpha)
-1	0.159
0	0.130
1	0.106
2	***0.085***
3	0.067
4	0.052
5	***0.040***
6	0.030
7	0.023
8	0.017
9	0.012
10	***0.008***
11	0.006
12	0.004
13	0.003
14	0.002
15	***0.001***
Deceptive (SR) Cut-scores	
Total SR Cut-score	p-value (alpha)
1	0.159
0	0.127
-1	***0.099***
-2	0.077
-3	0.058
-4	***0.043***
-5	***0.032***
-6	0.023
-7	***0.016***
-8	0.011
-9	***0.008***
-10	0.005
-11	0.003
-12	***0.002***
-13	***0.001***