Reliability of the Empirical Scoring System with Expert Examiners

Raymond Nelson, Benjamin Blalock, Marty Oelrich, and Barry Cushman

Abstract

Monte Carlo statistical methods were used to calculate confidence intervals for the reliability and accuracy of the ESS with a cohort of 25 experienced examiners who scored a small sample of 10 confirmed psychophysiological detection of deception (PDD, polygraph) exams that were conducted using the Federal ZCT format. Fleiss kappa showed a substantial agreement between the numerical scores of the study participants (k = .61, .54 to .68) and decision agreement that was significantly better than chance with a mean rate of decision agreement of 95.4%, excluding inconclusive results. Bootstrap Monte Carlo methods were used to calculate the accuracy profile and statistical confidence intervals of the ESS scores from the experienced examiners. The authors recommend continued interest in the ESS as an evidence based model for manual test data analysis in field polygraph settings and future research.

Introduction

This study is a small scale survey of inter-scorer reliability accuracy for the Empirical Scoring System (ESS) (Blalock et al., 2009; Krapohl, 2010; Nelson, Krapohl & Senter, 2008) with a cohort of experienced examiners (N = 25). The ESS is an evidence-based model for manual test data analysis (TDA) of psychophysiological detection of deception (PDD) exams, and is premised on a requirement for empirical support for all of the assumptions and procedures incorporated into the scoring model. The ESS was designed to provide field examiners and consumers of polygraph results with a simple statistical model to calculate the probability of an erroneous test result using hypothesis testing models based on normative data. The ESS model is capable of providing mathematical estimates of test sensitivity, test specificity, inconclusive rates and error rates using normative inferential statistical methods that are more resistant to unknown base rates or prior probabilities than previous TDA models that lack normative data and rely primarily on non-resistant Bayesian conditional calculations of accuracy and error rates. Unlike other manual TDA models, the ESS is based on normative data, and provides an inferential calculation of level of statistical significance of a manual score, including potential error and inconclusive rates, and known rates of inter-scorer reliability.

Previous studies on the ESS with inexperienced scorers (Blalock et al., 2009; Nelson et al., 2008) have shown that this simplified evidence-based model for TDA is capable of providing decision accuracy levels, including sensitivity, specificity, inconclusive, and error rates that equal or exceed those of existing complex scoring methods used by experienced examiners. Previous studies have also shown the ESS to provide inter-scorer reliability among inexperienced examiners that equaled or exceeded the reliability of previous scoring methods with experienced examiners. Krapohl (2010) showed that the ESS scores could be approximated from decremented 7-position scores, and that decision accuracy and agreement was essentially equivalent (ie, no statistically significant differences) to the results obtained from the 7-position scores. The present study was intended to further investigate the reliability and accuracy of the ESS with experienced examiners.

Method

Participants

A cohort of experienced PDD examiners (N = 25) participated in this study at a two-day continuing education seminar during which the ESS was taught for approximately three hours. Nineteen of the study participants were government examiners, three worked for civilian law

enforcement agencies, and three were employed in private practice. Experience ranged from 1 to 34 years in field practice (mean = 9.3 years, median = 5.0 years). The number of completed examinations ranged from 30 to 6000 (mean = 1846, median = 1300). Ten of the examiners were male; 15 were female. Nine of the participants had completed graduate level education; 13 scorers had undergraduate level education, and three others reported some college education. Ages of the participants ranged from 25 to 65 years (mean = 38 years, median = 37 years). One participant did not provide any age data. Seventeen of the participants completed their PDD training with the United States Defense Academy for Credibility Assessment (DACA, now the National Center for Credibility Assessment, NCCA), and eight participants were trained at polygraph schools accredited by the American Polygraph Association (APA). Twenty-one of the study participants reported using the manual TDA model taught at NCCA; one participant reported using the Backster TDA model; one participant reported using the Utah TDA model, one participant reported using the TDA model published by ASTM International (ASTM, 2002), and one participant did not report the TDA model generally used. Participation in the study was voluntary and had no effect on the employment, certification, or professional status of the participants.

There were no incentives, rewards or consequences of any kind associated with anyone's participation or performance during the study. Study participants completed the scoring tasks independently, and received no feedback on their performance. Study participants were provided with documentation of their attendance at the continuing education seminar, and all of the conference attendees participated in this voluntary experiment.

Apparatus

Printed paper charts were distributed to the participants who were instructed to record ESS scores on paper score-sheets. ESS scores were obtained by means of visual analysis of the PDD data with no mechanical or computerized measurements. Scores were later entered into a computer spreadsheet and analysed with a commercially available application, using statistical tools designed to calculate bootstrap, Monte Carlo, and inferential models.

Design

The cohort of 25 experienced polygraph examiners used the ESS to manually score a sample of 10 cases each, which were randomly selected from a larger sample of 100 confirmed Federal Zone Comparison Technique (ZCT; Department of Defense Polygraph Institute, 2006) (single-issue/event-specific) examinations, previously described and used in Krapohl & Cushman (2006). Deceptive cases were determined by the examinees' confession or other substantial evidence, while truthful cases were confirmed through the confession or substantial evidence of guilt for someone other than the examinee. No effort was made to match truthful and deceptive cases during random selection, and the participants scored an unequal number of truthful and deceptive cases, four and six, respectively.

All examinations consisted of three relevant questions and three charts. Two-hundred fifty categorical decisions were made by the 25 participants who were asked to score the 10 cases individually, without assistance from others. One-hundred categorical decisions were made on the truthful cases 150 decisions deceptive cases. Numerical scores were obtained for 750 investigation target questions, including 300 confirmed truthful questions and 450 confirmed deceptive test questions. Overall, 2350 individual component scores were provided by the study participants, including 900 component scores for confirmed truthful test questions and 1,250 component scores for confirmed deceptive test questions. One-third of the component scores referred to each component sensor (ie, pneumograph, electrodermal [EDA], and cardiograph). Categorical decisions and scores for the test questions and component sensors were analysed using a series of bootstrapping and Monte Carlo experiments. No feedback or review was provided during the data collection, and the participants were given no extra-polygraphic information regarding the examinee, case matter, case status or outcome.

Empirical Scoring System scoring criteria are based on the primary physiological

features described in multiple studies on PDD feature extraction and feature development (cf. Harris et al., 2000; Kircher et al., 2005; Kircher & Raskin, 1988; Raskin et al., 1988). Scores are assigned using a non-parametric three-position transformation rubric (Van Herk, 1990; Department of Defense Polygraph Institute, 2006) based on the seven-position scoring model (Backster, 1963) in which integer scores are assigned based on the evaluation of the strength of reaction to target and comparison stimuli. The ESS uses a the *bigger-is-better* rule (Department of Defense Polygraph Institute, 2006) and makes no assumptions about the linearity of physiological response data (Handler, Nelson Krapohl & Honts, 2010). Because several studies have suggested that the EDA data is the strongest contributor to the final score and generally accounts for approximately half of the final score of all examinations (cf. Capps & Ansley, 1992; Harris et al., 2000; Kircher & Raskin, 1988; Krapohl & McManus, 1999; Kircher et al., 2005; Nelson et al., 2008; Raskin et al., 1988), all EDA scores were doubled, regardless of the magnitude of differential response observed between the target and comparison stimuli. Like other manual scoring models, ESS numerical scores are summed to achieve a grand-total score for the test as a whole, and sub-total scores for the individual target questions.

Interpretation of the ESS numerical scores (ie, translation into usable human language), and classification of test results were based on two-stage decision rules. Two-stage decision rules (Krapohl, 2005; Krapohl & Cushman, 2006; Senter, 2003; Senter & Dollins, 2008) provide reduced inconclusives and increased sensitivity to deception compared to the grand total rule. Two-stage decision rules prioritize the importance of the grand-total score at stage one, and employ the sub-total scores during stage two only when the grand-total score is inconclusive during stage one. In contrast, other decision rules, which allow the sub-total score to supersede the grand-total score, or which place arbitrary (ie., non-evidence based) requirements on the numerical sign value of sub-total scores, have been shown to produce elevated rates of inconclusive and false positive errors among truthful examinees (Krapohl, 2005; Krapohl & Cushman, 2006; Senter, 2003; Senter & Dollins, 2008).

Krapohl (1998) demonstrated that adjustment of decision cut scores could alter the sensitivity, specificity, inconclusive and error rates of a manual scoring model. However, little work has been done to develop normative data for most manual scoring models that are presently used in field settings. ESS decision cut scores were selected using Monte Carlo norms and statistical principles that are inherent to hypothesis testing and the scientific method. These principles involve a declaration of an alpha level prior to testing or scoring, representative of a desired level of decision accuracy or maximum tolerable proportion of errors. Error boundaries for previous studies on the ESS (Blalock et al., 2009; Krapohl, 2010; Nelson et al., 2008) used (alpha = .1) for truthful classifications and (alpha = .05) for deceptive classifications. Normative data for the selection of numerical cut scores that correspond to the specified alpha levels were reported by Nelson et al. (2008).

Any use of sub-total scores and decision rules that make use of the individual target questions in a polygraph examination of a single known or alleged incident will result in a condition in which multiple statistical comparisons are performed on the single known or alleged incident. This will result in a well-known potential statistical problem known as *inflated alpha*, in which the stated tolerance rate for type-1 error rate is compounded by the number of statistical comparisons (ie, target questions). In polygraph testing, this means that the addition of relevant questions to the test, combined with the spot score rule, will increase false positive errors unless the effects of the additional questions are accommodated statistically. To avert the predictable increase in false-positive errors that results from the inflation of alpha, we employed a Bonferonni correction to the desired alpha (.05) by dividing it by the number of statistical comparisons that were to be completed. In the case of ZCT PDD exams there are three relevant questions. Therefore a Bonferonni corrected alpha of .0167 (3 * .05 = 0.0167) was used to make deceptive classifications on the three sub-total scores.

Analysis

Inter-scorer reliability of the categorical results of the 25 experienced

examiners was the primary purpose of this study. Fleiss' kappa was calculated as a measurement of inter-rater agreement, and bootstrap methods were used to calculate empirical values for the 95% confidence intervals for the kappa statistic. In addition, bootstrap means and confidence intervals were calculated using 1,000 resampled sets of 600 pairwise comparisons of decisions made by the 25 study participants, excluding pairs in which one or more of the participants achieve an inconclusive result.

Of lesser interest, due to the small number of cases which the study participants scored, was the decision accuracy and inconclusive rates obtained by the study participants. Bootstrap Monte Carlo methods were used to calculate a dimensional profile of accuracy, and empirical confidence intervals, achieved by 25 study participants, including: percent correct; inconclusive results for deceptive, truthful, and combined groups; sensitivity to deceptive; specificity to truthfulness; false negative errors; false positive errors; positive predictive value (PPV); negative predictive value (NPV);[1] proportion of correct decisions for deceptive cases; proportion of correct decisions for truthful cases; and the unweighted average of correct decisions for the deceptive and truthful groups. Poisson analysis was used to evaluate the statistical significance of errors observed in the scores provided by the study participants.

Results

The 25 study participants produced a Fleiss' kappa score of (K = .61), which is indicative of a substantial level of inter-scorer agreement according to the interpretation scheme suggested by Landis and Koch (1977)[2]. Table 1 shows the kappa statistics along with similar calculations from previous

Table 1. Fleiss' kappa reliability statistics and (95%) confidence intervals

Sample	Kappa (95% confidence interval)
Experienced examiners	.61 (.54 to .68)
Inexperienced examiners from Nelson *et al.*, 2008	.61 (.52 to .69)
Experienced examiners from Nelson *et al.*. 2008	.57 (.50 to .65)
Inexperienced examiners from Blalock *et al.*. 2009	.56 (.48 to .63)
Blackwell (1999) - experienced federal examiners	.57 (none reported)

[1] PPV and NPV are calculated from the combined groups of confirmed truthful and confirmed deceptive cases. PPV is the conditional probability that a positive (i.e., deceptive) is correct, and is calculated as the ratio of true-positives to all positives (i.e., true-positives and false-positives). NPV is the conditional probability that a truthful result is correct, and is calculated as the ratio of true-negatives to all negatives, including true-negatives and false-negatives. PPV and NPV are referred to as *conditional* probabilities because they are non-resistant to (i.e., affected by) the prior probability of deception or truthfulness. Prior probabilities are estimated by various methods including external evidence and known or estimated base-rates. Observed PPV and NPV will change with differences in base-rates. In contrast, test sensitivity and test specificity rates are resistant to base rates, along with normative calculations of statistical significance or probabilities of error. Polygraph accuracy profiles will be best informed by a complete profile that includes all dimensional aspects of test accuracy including both resistant and non-resistant dimensions.

[2] Landis and Koch (1977) suggested the following interpretation scheme for Fleiss' kappa values: <0 = poor agreement, 0.00 to 0.20 = slight agreement, 0.21 to 0.40 = fair agreement, 0.41 to 0.60 = moderate agreement, .61 to .80 = substantial agreement, .81 to 1.00 = near perfect agreement.

studies. Also included in Table 1 are bootstrap calculations of statistical confidence intervals for the kappa statistics.

A second empirical bootstrap of 1000 resampled sets of the proportion of agreement between decisions produced by the 25 participants showed the bootstrap mean rate of agreement between decisions made by the 25 study participants was of 95.4% (SEM = 93.3% to 97.2%). The bootstrap range of all agreement scores was from 81.4% to 100%.

Although the present study is intended primarily as an investigation of reliability and decision agreement for experienced examiners and is not intended to be an investigation of decision accuracy for the polygraph or the ESS model, readers may be interested in the decision accuracy achieved by the experienced examiners who participated in this study. Decision accuracy with inconclusives was .867 for the combined deceptive and truthful cases. Decision accuracy for deceptive cases alone was .935 among the 25 study participants, and 1.000 for truthful cases, when inconclusive results were removed. The rate of inconclusive results was .108 for all cases, including .139 for deceptive cases and .063 for truthful cases. False negative errors were observed at .042, and there were no false-positive errors made by the study participants.

Poisson analysis, used to calculate the probability of the occurrence or frequency of rare or unusual events, was used to investigate the absence of false-positives. Krapohl (2006) reported a false-positive error rate of 14.0% among truthful cases for the studies cited by Blackwell (1998), Krapohl (2005), and Yankee, Powell, & Newland (1985), and along with an inconclusive rate of 23.0% among truthful cases, for Federal ZCT exams. Decision accuracy was reported as 82.0% for the studies cited. For deceptive cases, Krapohl (2006) reported an inconclusive rate of 9.0%, decision accuracy at 97.0%, and false-negative errors at a rate of 2.7%. Using the 14.0% false-positive rate as the expected mean frequency, Poisson analysis indicates a high probability of observing zero false-positive errors due to random chance alone (p = .57) in a confirmed case sample of the size used in the present study. Therefore, the false-positive error rate and positive predictive value observed during this study is statistically meaningless and not generalizable to field settings. Readers should be cautioned against any expectation to never encounter false-positive errors in larger studies or in field settings.

To further investigate the level of criterion accuracy achieve by the 25 study participants, bootstrap Monte Carlo methods were used to calculate an accuracy profile and confidence intervals, using the study participant scores as seed values. Table 2 shows the Monte Carlo accuracy profile for the experienced examiners in the present study.

Discussion

All scientific studies are fraught with some limitations. The present study is limited most obviously by the small size of the confirmed case sample that was scored by the study participants. Despite the sample size limitation, the present study does include a moderately large number of participants, compared to other PDD studies. The total number of question scores and examination scores provides a large database of seed values for statistical analysis of inter-scorer reliability among experienced examiners. While a cohort of experienced examiners who voluntarily participate in a test data analysis experiment, in the context of their attendance at a continuing education seminar, may not be representative of all field examiners, it does suggest that the ESS can generalize to and be used effectively by more experienced PDD examiners. Previous studies on the ESS have shown the effectiveness of the ESS with inexperienced scorers. Coupled with powerful statistical models, the results from this small study are not precluded from providing some useful information. Of course, no single study is capable of providing definitive answers to research questions. Results from this study should be regarded only as preliminary findings from a pilot study.

With the study limitations in view, the results of this reliability survey among experienced examiners are interesting, and concur with the results of earlier studies using inexperienced examiners. The ESS model provides a high level of reliability when

Table 2. Bootstrap Monte Carlo accuracy profile and empirical confidence intervals.

Dimension	Proportion (95% CI)
Correct	.956 (.681 to>.999)
INC	.102 (.010 to .355)
T INC	.083 (<.001 to .217)
D INC	.121 (<.001 to .521)
Sensitivity	.866 (.609 to>.999)
Specificity	.860 (.269 to>.999)
FP	.049 (<.001 to .200)
FN	.018 (<.001 to .180)
PPV	.977 (.768 to>.999)
NPV	.934 (.545 to>.999)
D Correct	.943 (.756 to>.999)
T Correct	.968 (.583 to>.999)
Unweighted accuracy	.955 (.655 to>.999)

manually scoring PDD data, and provides an expedient and effective model for manual test data analysis of PPD exams. The ESS is easy to learn, and its simplicity can be expected to contribute to field practitioner skills that are highly reliable for both experienced and inexperienced examiners. In field settings, simplified test data analytic models based in published scientific evidence will be less perishable than skills based on many complex and unproven rules and complex un-studied assumptions. In addition, simpler decision and classification models may be less subject to over-fitting a development sample, and may generalize to field settings. Simple evidence-based decision models may also be less subject to modification by creative or forgetful field practitioners, when compared with complex decision models based on theory or expert opinion alone. In research settings, the structure and simplicity of the ESS will mean that PDD researchers who are not field examiners can have access to a reliable method for manual test data analysis. Additional studies should investigate the ability of non-polygraph professionals to learn and use the ESS when reviewing or analyzing the results of PDD examinations, and the possibility of automating the ESS procedures.

The criterion validity levels observed during this study are impressive, and the experienced examiners using the ESS produced an accuracy profile that generally outperformed those of inexperienced examiners from previous studies on the ESS. However, these study results should not be generalized as an indicator of criterion validity

levels without additional support in the form of larger experimental investigations of any differences in criterion validity among experienced and inexperienced scorers.

Coupled with the results of powerful computer scoring algorithms, for which their functions and validity are completely documented and available for use and expert review, the ESS model provides the field examiner with the capability to answer the challenging questions that will undoubtedly be raised by scientific minded opponents and challengers of PDD testing programs. Those questions will pertain to the reliability levels, test sensitivity, test specificity, inconclusive rates, false-positive and false-negative error rates, and the level of statistical significance or probability of error when classifying a test result as either truthful or deceptive.

It is important to note that the ESS is not based in new methods or new knowledge. It is founded on decades of research by well-established professional investigators of the scientific foundations of PDD testing. Feature development studies that led to the ESS include research conducted by Harris et al., (2000), Kircher et al., (2005), Kircher & Raskin, (1988), and Raskin et al., (1988). Transformation methods for the ESS have their origins with Backster (1963), and others including Bell, Raskin, Honts & Kircher (1999), Blackwell (1998), Harwell (2000), Krapohl (1998), and Van Herk, (1990). Barland (1985) proposed the equivariance Gaussian signal discrimination model as a basis for polygraph decisions. Decision rules employed in the ESS are the product of studies conducted by Senter (2003) and Senter & Dollins (2008), and the selection of statistically optimal cut scores based on normative data has emerged from the work of Krapohl (1998, 2002), Krapohl & McManus (1999), and Nelson et al. (2008).

It should go without saying that all scientific studies of polygraph decision models and scoring accuracy, both manual and automated, should provide a complete structural description of the decision model and supporting scientific data, including empirical evidence on feature development, transformations and mathematical models, decision rules, and normative data that can be used to calculate and evaluate model effectiveness in terms of sensitivity, specificity, inconclusive rates, error rates and inter-rater reliability Moreover, all calculations based on sample data, and all results from scientific study should be regarded as estimates of actual field performance.

It is a common expectation that test and model developers provide, in publication, evidence of statistical power, in the form of calculations of statistical power, calculations of statistical errors of measure, variance estimates, or statistical confidence intervals that will inform the scientifically minded reader of the strengths and limitations of new knowledge derived from a study. The ESS is capable of addressing all of these concerns in an evidence-based manner, and is supported by numerous published studies, by a variety of researchers and institutions, on the development and validation of the structural components that make up the ESS model.

In the absence of any evidence against the effectiveness of the ESS with both experienced and inexperienced scorers, additional research interest and validation studies on the ESS are recommended. While the present study is based on a very small case sample, it is not completely uninformative. Larger scale experiments and replication studies should continue to investigate the ESS with experienced examiners, inexperienced scorers, non-polygraph professionals, and automated models. Field examiners and program managers who are looking for a reliable, validated, evidence-based scientific approach to PDD test data analysis should be encouraged to consider the many empirical and practical advantages of the ESS model in field settings.

References

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

Backster, C. (1963). Polygraph professionalization through technique standardization. *Law and Order*, 11, pp. 63-65.

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

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

Blackwell, J. N. (1998). *PolyScore 33 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations*. 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, 281-288.

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

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.

Handler, M., Nelson, R., Goodson, W., & Hicks, M. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph*, 39(4), 200-215.

Handler, M., Nelson, R., Krapohl, J. & Honts, C. (2010). An EDA primer for polygraph examiners. *Polygraph*, 39, 68-108.

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.

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

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.

Kircher, J. C. & Raskin, D.C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, 73, 291-302.

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

Krapohl, D. J. (2002). Short report: 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. (2006). Validated polygraph techniques. *Polygraph*, 35(3), 149-155.

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

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.

Landis, J. R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

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

Raskin, D. 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.

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

Senter, S. M. & Dollins, A.B. (2008). Exploration of a two-stage approach. *Polygraph*, 37(2), 149-164.

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.

Yankee, W. J., Powell, J. M, III & Newland, R. (1985). An investigation of the accuracy and consistency of polygraph chart interpretation by inexperienced and experienced examiners. *Polygraph*, 14, 108-117.