Criterion Validity of the Empirical Scoring System with Experienced Examiners: Comparison with the Seven-Position Evidentiary Model Using the Federal Zone Comparison Technique

Raymond Nelson and Donald Krapohl

Abstract

This study is a comparison of criterion accuracy profiles of psychophysiological detection of deception examinations when manually scored by a cohort of experienced federally trained examiners who evaluated a confirmed case sample (N = 60) of Federal ZCT examinations. More than 50% of subtotal (spot) scores among the confirmed truthful cases were non-positive values. Seven-position scores were to Empirical Scoring System (ESS) scores using two-stage decision rules. There were no statistically significant differences between criterion accuracy of the ESS and the seven-position evidentiary model within the profiles of 13 dimensions of criterion accuracy. The authors recommend continued interest in the ESS and seven-position models, and make recommendations for further research.

Introduction

The Empirical Scoring System (ESS) is an evidence-based model for manual test data analysis (TDA) of psychophysiological detection of deception (PDD) examination data (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Krapohl, 2010; Nelson, Blalock, Oelrich & Cushman, 2011; Nelson, Krapohl & Handler, 2008). The ESS is a three-position model for TDA (Harwell, 2000; Krapohl, 1998; Van Herk, 1990), which is a modification of the seven-position model (ASTM, 2002; Backster, 1963; Bell, Raskin, Honts & Kircher, 1999; Department of Defense Polygraph Institute, 2006) taught at the National Center for Credibility Assessment and other polygraph schools accredited by the American Polygraph Association. The ESS uses a weighted transformation of electrodermal scores, while the seven-position model achieves component weighting through the distribution of scores within the seven-position scale. Both the ESS and the seven-position scoring model are based on physiological features that were identified as valid by scientists at the University of Utah (Kircher & Raskin, 1988; Kircher & Raskin, 2002; Kircher, Kristjansson, Gardner & Webb, 2005; Raskin, Kircher, Honts & Horowitz, 1988).

A primary decision rule for the ESS, the two-stage rule (TSR) or “Senter rule” was developed at the Department of Defense (Senter, 2003; Senter & Dollins, 2002; Senter & Dollins, 2008a; Senter & Dollins, 2008b). Cutscores for the traditional seven-position model are those that have been historically taught at accredited polygraph schools, but for which normative data have never been published. Cutscores for the ESS are based on normative data (Nelson, et al., 2008) and are selected for a desired level of statistical significance.

Development and validation of the ESS has primarily focused on model effectiveness and criterion validity with inexperienced scorers. A principal goal was to simplify the scoring process, a step considered beneficial to inter-scorer reliability and to training requirements. A second equally important goal was to ensure each step in the scoring system maximized decision accuracy.

One study (Nelson et al., 2011) showed the ESS to provide high levels of decision agreement and criterion accuracy using the ESS. Criterion validity of the seven-position model has been described in several studies, and was summarized by Krapohl, (2006). However, the studies cited by Krapohl (2006) included scores using an older seven-position model for TDA, for which the physiological features were replaced during 2006 (Department of Defense Polygraph Institute, 2006), while leaving cutscores and decision rules unchanged from the older seven-position federal TDA model (Light, 1999; Swinford, 1999). At the present time there are no published studies that describe the criterion validity of the current seven-position model for TDA with the Federal Zone of Comparison Technique (ZCT). This study is an experimental comparison of ESS and seven-position scores obtained from experienced federally trained examiners.

Method

Data

Six experienced federally trained examiners scored a sample of confirmed PDD examinations using the seven-position scoring system (DoDPI, 2006). The examinations had been conducted using the Federal ZCT, an event-specific PDD technique that is widely taught and considered to be among the most accurate of all diagnostic PDD techniques. All examinations consisted of three relevant questions, three probable-lie exclusionary comparison questions, and three test charts. Thirty cases were confirmed as truthful, the remaining 30 as deceptive. Confirmation for all cases was obtained in the form of extrapolygraphic evidence such as physical evidence of guilt or innocence, physical evidence of guilt of an alternative suspect, or the confession of an alternative suspect.

Cases were divided into six subsets, with 10 cases in each subset using a sequence of random numbers. Subsets were then randomly assigned to the six study participants who remained blind to the ground truth status of each case. Each participant scored 10 cases. Variation among scorers is assumed to improve the generalizability of the study results compared to results obtained from a single scorer.

Once the seven-position scoring was completed, the scores were transformed to ESS scores by collapsing the seven-position values to their three-position equivalents and then doubling all electrodermal scores. This provided us with two sets of scores for the same cases: ESS and the traditional seven-position scores. The scores in both sets were summed for each relevant question (subtotal) and for the entire case.

Evaluation of truthful cases revealed that 75.5% (SD = 8.4%) [95% CI = 59.0% to 91.9%] of the truthful cases had non-positive subtotal scores using ESS scores, and 78.9% (SD = 7.9%) [95% CI = 63.3% to 94.5%] of the subtotal scores were non-positive for truthful cases with the seven-position scores. A heteroscedastic t-test of the difference between the rates of non-positive confirmed truthful subtotal scores for the ESS and seven-position models was not statistically significant (*p* = .33). Decision rules that rely on all positive subtotals as a condition of making a decision of truthfulness would produce very low accuracy with truthful cases in this sample.

To reduce the influence of non-positive subtotal scores on this experimental comparison, all decisions for ESS and seven-position scores were made using two-stage rules (Senter & Dollins, 2002; Senter, 2003), in which the grand-total score always supersedes the subtotals scores. The TSR is the standard approach with ESS, but not with the federal decision rules. A related method called evidentiary decision rules (Krapohl, 2005; Krapohl & Cushman 2006) was designed to employ both federal scoring and TSR. It was applied to the federal scores to allow equivalency of data treatment, and a comparison of the scoring regimens of ESS and the federal system to assess the diagnosticity of the scores themselves.

Evidentiary rules begin with the total score, for which, if it meets or exceeds the decision thresholds, a definitive decision of Deception Indicated (DI) or No Deception Indicated (NDI) is made. Only when the total score result is inconclusive is the second stage invoked. The second stage is based only on the subtotals, of which there was one for each of the three relevant questions in the federal ZCT cases used in this study. If any of the subtotals fall at or below the required cutscore, the results would be DI instead of inconclusive. All other decisions would be inconclusive.

Decision Rules

Evidentiary decision rules for the seven-position scoring system are as follows:

1. If the grand total is +4 or greater, a decision of NDI is made.

2. If the grand total of all scores is -6 or lower, a decision of DI is made.

3. If the grand total is between +3 and -5, inclusive, the second stage is used. In this step, when a subtotal score is -3 or lower, a decision of DI is made.

4. All other cases would be called inconclusive (Inc).[1]

One of the strengths of the ESS is that decision rules can be established based on tolerance for error and inconclusives. In this study we tested two sets of decision thresholds: one set that produced an alpha of .10 for truthfulness and .05 for deception (labeled TSR .10 .05), and a second, more risk aversive set (labeled TSR .05 .05), which set alpha at .05 for both deception and truthfulness.

Two-stage decision rules with statistically optimal cutscores for the TSR .10 .05 model are as follows:

1. If the grand total is +2 or greater, a decision of NDI is made.

2. If the grand total of all scores is -4 or lower, a decision of DI is made.

3. If a subtotal score is -7 or lower, a decision of DI is made.

4. All other cases would be called inconclusive.

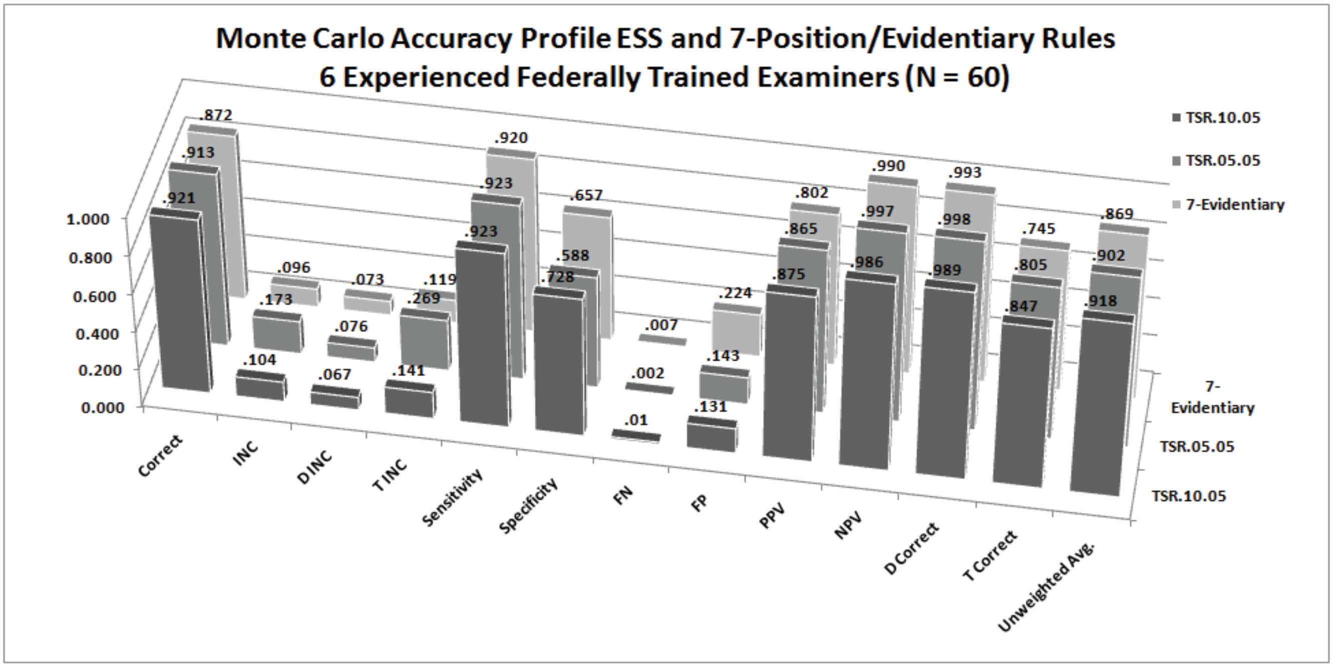
Two-stage decision rules with statistically optimal cutscores for the TSR .05 .05 model are as follows:

1. If the grand total is +5 or greater, a decision of NDI is made.

2. If the grand total of all scores is -4 or lower, a decision of DI is made.

3. If a subtotal score is -7 or lower, a decision of DI is made.

4. All other cases would be called inconclusive.



[1] In field practice, if the result would be inconclusive based on the numerical scores, examiners can collect up to two additional test charts to garner more data. The scores are then summed for all charts, and the same decision rules would apply. Because the cases in the present study were conducted some years prior to the implementation of the practice of conducting additional testing, the inconclusive rate reported here may not accurately represent the current rate.

Results

Results were evaluated along several indices of criterion accuracy, including: overall decision accuracy, total inconclusives, inconclusive truthful cases, inconclusive deceptive cases, sensitivity to deception, specificity to truthfulness, false-negative rate, false-positive rate, positive predictive value, negative predictive value, percent correct for deceptive cases, percent correct for truthful cases, and the unweighted average of the percents correct for truthful and deceptive cases. Mean scores and statistical confidence intervals were calculated using Bootstrap Monte Carlo methods, along with standard deviation scores that were used to calculate variance and sums of squares that were then used to calculate a planned series of one-way ANOVAs (See Table). There were no statistically significant differences observed in any of the dimensional characteristics of criterion accuracy between the ESS and the seven-position model with evidentiary decision rules.

Table. Monte Carlo accuracy profiles for ESS and seven-position scores

Mean (Standard Deviation) [95% Confidence Interval]				
N = 60	ESS TSR.10.05	ESS TSR.05.05	7-Position Evidentiary	ANOVA (df = 2,177) F (p)
Correct	.921 (.037) [.849 to .993]	.913 (.414) [.834 to .991]	.872 (.045) [.783 to .961]	0.707 (.495)
INC	.104 (.039) [.027 to .181]	.173 (.058) [.079 to .267]	.096 (.038) [.022 to .171]	1.393 (.251)
D INC	.067 (.046) [.001 to .158]	.076 (.427) [.001 to .172]	.073 (.049) [.001 to .169]	0.014 (.986)
T INC	.141 (.063) [.017 to .264]	.269 (.037) [.113 to .426]	.119 (.060) [.002 to .236]	1.919 (.150)
Sensitivity	.923 (.049) [.826 to 1.02]	.923 (.500) [.826 to .999]	.920 (.050) [.821 to .999]	0.002 (.998)
Specificity	.728 (.082) [.568 to .889]	.588 (.051) [.412 to .764]	.657 (.088) [.484 to .829]	1.001 (.370)
FN	.010 (.019) [.001 to .047]	.002 (.268) [.001 to .015]	.007 (.015) [.001 to .036]	0.145 (.865)
FP	.131 (.062) [.009 to .253]	.143 (.428) [.016 to .269]	.224 (.076) [.076 to .373]	0.962 (.384)
PPV	.875 (.059) [.759 to .991]	.865 (.437) [.746 to .984]	.802 (.069) [.668 to .937]	0.643 (.527)
NPV	.986 (.025) [.937 to 1.035]	.997 (.286) [.974 to 1.021]	.990 (.023) [.946 to .999]	0.124 (.884)
D Correct	.989 (.020) [.950 to .999]	.998 (.257) [.984 to .999]	.993 (.016) [.961 to .999]	0.144 (.866)
T Correct	.847 (.072) [.707 to .988]	.805 (.002) [.635 to .974]	.745 (.085) [.578 to .912]	0.626 (.536)
Unweighted Avg.	.918 (.037) [.846 to .991]	.902 (.008) [.816 to .987]	.869 (.043) [.784 to .954]	0.578 (.562)

Discussion

Previous studies have shown high levels of criterion validity of ESS scores with inexperienced scorers. The results of this study show a high level of criterion validity for both the ESS and the seven-position evidentiary models with experienced scorers. (For an accuracy comparison between the seven-position evidentiary decision rules and the standard federal decision rules, see Krapohl, 2005; Krapohl & Cushman, 2006.) Although the ESS showed marginal increases in criterion accuracy, in terms of raw frequencies along some indices, the differences were not statistically significant. Holding the decision rules constant for the ESS and seven-position model means that any observed difference would be due to the numerical scores and the cutscores, which are a feature of the normative data. These results suggest that it is possible that observed differences in criterion validity may go unnoticed by experienced field examiners who use evidentiary decision rules with the seven-position scoring model, compared to results that would be achieved with the ESS model.

Limitations of the present study include the small cohort of scorers, a relatively small sample size, and little access to information about how the confirmed cases came to be included in the federal archive. Another limitation of the present study is that the 7-position scores from experienced scorers may not be representative of 7-position scores obtained from inexperienced scorers. No effort should be made to represent a single sample, single cohort, or single study as indicative of the definitive answer regarding complex questions of model validity. Another limitation of the present study is that the accuracy profile produced by the seven-position evidentiary model can be expected to differ from that of the seven-position model using traditional ZCT decision rules, which were not included in the present analysis. The absence of any statistically significant difference in criterion accuracy in this study suggests that the selection of a scoring model may ultimately be a matter of administrative or field practice policy, which should be made according to operational priorities.

Although the numerical scores themselves do not seem to contribute to statistically significant differences in criterion validity profiles for the ESS and seven-position models, it is possible that different decision rules would produce statistically significant differences in some dimensions of criterion accuracy. Additionally, while there are no significant differences in criterion validity between the ESS and seven-position evidentiary models, some operational differences may be experienced by field examiners using the ESS and seven-position models due to the increased complexity of the seven-position model and the simplicity of the ESS. Simpler models may provide some advantages in terms of inter-scorer agreement, skill acquisition, skill retention, and generalizability. This difference may be more important to less experienced examiners, and should be the focus of future research. In addition, future research should further evaluate the contribution of decision rules and cutscores to the accuracy profiles achieved by both experienced and inexperienced scorers using the ESS and seven-position models. The rate of non-positive subtotal scores and variability between subtotal scores among confirmed truthful cases suggests that further research may lead to decision rules that make optimal use of subtotal scores. Normative studies on seven-position scores should also be published, and statistically optimal cutscores should be investigated for the seven-position model. Continued interest in the ESS and seven-position models is recommended, in both research and field practice 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, 63-65.

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

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

Department of Defense Polygraph Institute (2006). Test Data Analysis: DoDPI numerical evaluation scoring system. Reprinted in *Polygraph*, 40(1).

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.

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

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

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.

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

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

Krapohl, D.J. (2010). Short Report: A Test of the ESS with Two-Question Field Cases. *Polygraph*, 39, 124-126.

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.

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

Nelson, R., Blalock, B., Oelrich, M. & Cushman, B. (in press). Reliability of the Empirical Scoring System with expert examiners. *Polygraph*, 40.

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). Validity of control question polygraph tests in criminal investigation. *Psychophysiology*, 25, 476.

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. Report number DoDPI00-R-0001. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC.

Senter, S. M. & Dollins, A.B. (2008a). Optimal decision rules for evaluating psychophysiological detection of deception data: an exploration. *Polygraph*, 37(2), 112-124.

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

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

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.