Monte Carlo Study of Criterion Validity of the Directed Lie Screening Test using the Seven-Position, Three-Position and Empirical Scoring Systems

Raymond Nelson

Abstract

Monte Carlo methods were used to calculate criterion accuracy profiles for DLST examinations scored with the seven-position, three-position, and ESS models. Multivariate ANOVAs were used to compare the results. Decision accuracy was significantly greater than chance, over 87%, for DLST exams scored with all three TDA models, and there were no significant differences in the unweighted accuracies achieved by the three scoring methods. The ESS model produced the fewest inconclusive results. Differences in inconclusive rates were significantly greater for the three-position model, and the difference was loaded on deceptive cases. Results suggest that the component weighting achieved by the seven-position and ESS models is more effective than the three-position model at extracting diagnostic information from the raw test data.

Introduction

The Directed Lie Screening Test (DLST), also known as the Test for Espionage and Sabotage (TES) was developed at the US Department of Defense (Research Division Staff, 1995a; 1995b) as a psychophysiological detection of deception (PDD) technique for use in security screening contexts. Handler, Nelson and Blalock (2008) described the use of the DLST in other PDD screening contexts such as public safety screening programs and post-conviction sex offender supervision and treatment programs. Like all screening tests, the DLST is conducted in the absence of any known incident, known allegation, or known problem. Like other PDD screening formats, the DLST is designed for use with multiple independent[1] targets for which it is conceivable that an examinee may be involved in one or more target behaviors while remaining un-involved in other investigation targets.

The DLST is similar to other PDD formats in its use of test questions, including the use of multiple presentations of a sequence of reviewed target questions, supported by carefully constructed operational definitions that describe the examinee's possible behavioral involvement in the issue or issues of concern. Also included in the DLST are comparison questions, intended to evoke a measurable response

Raymond Nelson is a research specialist with the Lafayette Instrument Company and an elected member of the APA Board of Directors. The views expressed in this work are those of the author and not the LIC or the APA. Mr. Nelson is a psychotherapist, polygraph field examiner, developer of the OSS-3 scoring algorithm, and is the author of several publications on various polygraph topics. For information contact raymond.nelson@gmail.com.

Acknowledgments: Special thanks to Mark Handler for reading an early draft of this manuscript, and to the reviewer who evaluated and critiqued the multivariate analysis.

[1] *Independence*, in scientific testing of this type, refers to the idea that the criterion status one issue does not affect the criterion status of other target issues. Although there is often debate about the level of independence of the criterion variance of test questions in both multi-issue and multi-facet contexts, use of the spot score rule represents a practical acknowledgment of the assumption of independence. When criterion independence is not assumed, diagnostic accuracy is maximized by evaluating the overall test result or grand total before evaluating the results of the individual targets.

from a truthful person, along with other procedural questions that are not subject to numerical scoring. The DLST differs from other PDD screening formats in that the DLST is conducted with several presentations of the test stimuli within a single test question sequence. The test questions sequence includes two relevant questions (RQs) that are repeated at least three times. In contrast, traditional PDD testing formats are constructed of several iterations of the test questions during three to five repetitions of the question sequence, stopping after each presentation of the sequence. The design of the DLST includes increased requirements for the proportion of non-artifacted and interpretable data, and also includes protocols for reducing the occurrence of inconclusive results. Although not unique to the DLST, this PDD format is always administered with directed-lie comparison questions.

Development studies on the TES/DLST were based on the seven-position manual test data analysis (TDA) method taught at the Department of Defense during the 1990s (Department of Defense, 2006). Those studies produced an unweighted accuracy rate of .833, along with an unweighted inconclusive rate of .081.[2] Subsequent studies by Nelson and Handler (2012) and Nelson, Handler and Morgan (2012) showed that DLST examinations can be interpreted using the Empirical Scoring System (ESS) (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Krapohl, 2010; Nelson, Blalock, Oelrich & Cushman, 2011; Nelson & Handler, 2010; Nelson & Krapohl, 2011; Nelson et al., 2011) and the Objective Scoring System version 3 (Nelson, Krapohl & Handler, 2008). ESS scores of DLST exams produced an unweighted accuracy level of .984 with an inconclusive rate of .098. This study was designed to extend our knowledge of criterion accuracy of the DLST with the three-position TDA and other TDA models. The hypothesis was that all three scoring models would produce accuracy rates that are significantly greater than chance.

Method

Bootstrap Monte Carlo methods were used to simulate DLST examinations scores and calculate the dimensional profile of criterion accuracy. Three different versions of the Monte Carlo model were created, with the seven-position, three-position, and ESS scores.

Archival data from a previous study by Krapohl (2005) were used to seed the subtotal scores of the Monte Carlo model. The seed data consisted of a matched sample of N = 100 examinations from the Department of Defense confirmed case archive. Fifty of the seed cases were confirmed truthful, and the other 50 were confirmed deceptive. Seven-position subtotal scores were transformed to their corresponding three-position values, and electrodermal scores were weighted to produce ESS scores. Because previous studies (Bell, Kircher & Bernhardt, 2008; Horowitz, Kircher, Honts & Raskin, 1997; Kircher et al., 2005) have suggested that the pneumograph data may not be diagnostic with DLC exams, pneumograph scores were not included in the calculation of seed parameters for the Monte Carlo space. The mean deceptive seven-position subtotal score was -2.418 (SD = 3.818) and the mean seven-position truthful subtotal was 2.653 (SD = 3.618). Three-position seed values had a mean deceptive subtotal score of -1.585 (SD = 2.382) and a mean truthful subtotal score of 1.719 (SD = 2.253). ESS seed values had a deceptive subtotal mean of -3.031 (SD = 5.104) and a mean truthful subtotal of 3.265 (SD = 3.935).

Subtotal scores for the seven-position, three-position, and the ESS TDA models were resampled to seed three different versions of a Monte Carlo space that consisted of 100 simulated examinations. Each of the simulated DLST exams in the Monte Carlo space consisted of two RQs, and all three Monte Carlo models were designed to repeat any examination using the same criterion status in the event of an inconclusive result.

[2] Accuracy levels for these studies were previously reported differently, while excluding some false-positive and inconclusive results. Refer to the study reports for more information.

The criterion states of the two independent RQs in the simulated DLST exams were set by comparing a random number to a fixed base rate of .293, which was calculated as the Šidàk correction (Abdi, 2007) of the desired base rate of .5. The criterion state for each exam in the Monte Carlo space was set to deceptive if the criterion of either or both RQs was deceptive, and the criterion state was set to truthful only when the criterion state of the two independent RQs was truthful.

Subtotals for simulated deceptive RQs were seeded by resampling from the archival subtotals scores of confirmed deceptive examinations, while simulated truthful subtotal scores were seeded by resampling from confirmed truthful examinations. Independence of subtotal scores within each simulated exam was ensured by randomly selecting subtotal scores from different examinations, including different exam (i.e., examination targets), different examiners, different examinees, and different positions in the test questions sequences of the cases in the archival data. The Monte Carlo space was recalculated for 10,000 iterations.

Cutscores and decision rules for seven-position and three-position scores were those specified by the Department of Defense (2006). All subtotals were required to be positive and the grand total score must equal or exceed four to be classified as No Significant Response (NSR) or truthful. Any examination with a subtotal of -3 or less, or a grand total of -4 or less, would be classified Significant Response (SR) or deceptive. Examinations meeting neither of those classifications would be classified as inconclusive.

The decision rule for the automated ESS model was the Spot Score Rule (SSR) (Light, 1999; Swinford, 1999) for which a deceptive classification was made if the absolute value of any subtotal score equaled or exceeded the subtotal cutscore corresponding to the desired alpha for deceptive decisions. Truthful classifications were made only if all subtotal cutscores equaled or exceeded the subtotal cutscores corresponding to the alpha for truthful decisions. Alpha was set at .05 for deceptive classifications and alpha = .1 for truthful classifications.

Bonferonni correction to the alpha cutscore for deceptive classifications was not used with the DLST examinations because the SSR is premised on the assumption that the criterion variance of individual questions is not affected by and does not affect the other questions.[3] Screening sensitivity was also increased by forgoing the use of Bonferonni correction. An inverse of the Šidàk correction for independent issues was used to correct for the deflation of alpha that occurs when calculating the normative probability that an examinee would produce a statistically significant truthful result to all investigation targets while lying to one or more of the independent issues.

ESS cutscores corresponding to these alpha levels were -3 and +1, using the normative data shown by Nelson, Handler and Morgan (2012). Any subtotal score of -3 or lower would be statistically significant for deception ($p < .05$), while test results in which all subtotal scores are +1 or greater would be statistically significant for truth-telling ($p < .1$).

Results

Alpha was set at .05 for all calculations of statistical significance.

Dimensional profiles of criterion accuracy were calculated, including means, standard deviations, and statistical confidence intervals for test sensitivity to deception, specificity to truthfulness, false-negative and false-positive errors, positive and negative predictive value, the proportion of correct decisions for deceptive and truthful cases excluding inconclusives, along with the unweighted average decision accuracy and unweighted inconclusive rates for the combined deceptive and truthful cases. Table 1 shows the results.

[3] It is sometimes the case that the behavioral details of the investigation target questions are not completely independent. However, use of the spot-score-rule is a practical acknowledgment of the assumption of criterion independence.

Table 1. Means, (standard errors) and {95% confidence intervals} for criterion accuracy

	7 Position	3 Position	ESS
Unweighted Accuracy	.874 (.034) {.807 to .941}	.893 (.030) {.833 to .953}	.871 (.035) {.803 to .939}
Unweighted INC	.096 (.03) {.038 to .154}	.208 (.039) {.131 to .284}	.047 (.021) {.006 to .089}
Sensitivity	.910 (.039) {.834 to .986}	.829 (.050) {.732 to .926}	.935 (.037) {.863 to .999}
Specificity	.677 (.041) {.597 to .757}	.595 (.041) {.515 to .675}	.730 (.041) {.650 to .810}
FN	.037 (.026) {.001 to .088}	.033 (.020) {.001 to .072}	.046 (.030) {.001 to .104}
FP	.184 (.054) {.078 to .289}	.127 (.043) {.042 to .212}	.195 (.056) {.085 to .306}
D INC	.053 (.032) {.001 to .115}	.138 (.047) {.047 to .229}	.020 (.019) {.001 to .058}
T INC	.139 (.050) {.041 to .237}	.277 (.061) {.158 to .397}	.075 (.038) {.001 to .149}
PPV	.830 (.051) {.730 to .930}	.884 (.044) {.798 to .970}	.824 (.050) {.726 to .921}
NPV	.949 (.038) {.875 to 1.023}	.966 (.030) {.907 to .999}	.942 (.038) {.867 to .999}
D Correct	.961 (.028) {.907 to 1.016}	.962 (.023) {.917 to .999}	.953 (.031) {.893 to 1.014}
T Correct	.786 (.061) {.667 to .906}	.824 (.056) {.714 to .934}	.789 (.060) {.671 to .907}

A 2 X 2 X 3-way ANOVA, criterion dimension (i.e., correct decisions and inconclusive results) x criterion state (i.e., deceptive or non-deceptive) x TDA model was used to compare the results of the seven-position, three-position, and ESS TDA models. Figure 1 shows the plots of means and 95% confidence intervals for decision accuracy and inconclusive rates for deceptive and truthful cases. Table 2 shows the three-way ANOVA summary. The three-way interaction of criterion dimension, case status and TDA model was significant. All two-way interactions and main effects were also significant in the three-way analysis. A series of post-hoc two-way and one-way ANOVAs were completed.

Figure 1. Means and 95% confidence intervals for correct decisions and inconclusive results of truthful and deceptive DLST cases scored with the seven-position, three-position, and ESS TDA models.

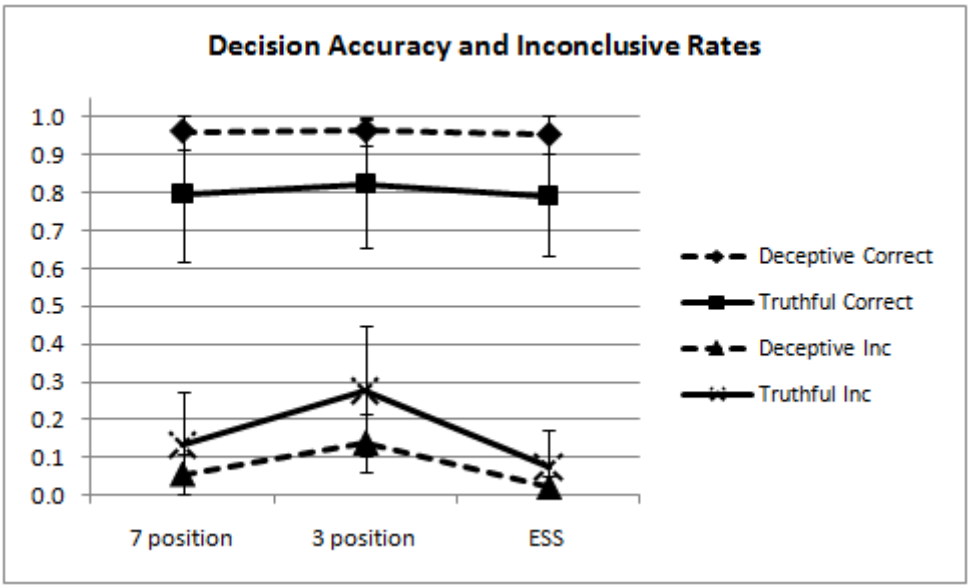


Table 2. Three-way ANOVA summary.

Source	SS	df	MS	F	p	F crit .05
Criterion dimension	87.631	1	87.631	25105.520	.000	3.857
Status	0.033	1	0.033	9.409	0.002	3.857
TDA Model	0.867	2	0.434	124.246	.000	3.011
Criterion dimension x Status	2.387	1	2.387	683.835	.000	3.857
Status x TDA Model	0.080	2	0.040	11.511	.000	3.011
Criterion dimension x TDA Model	0.507	2	0.254	72.641	.000	3.011
Criterion dimension x Status x TDA Model	0.023	2	0.012	3.334	.036	3.011
Error	2.052	588	0.003		.000	
Total	93.581	599			.000	

Two-way analysis of decision accuracy showed no significant interaction between TDA model and case status, and no significant main effect for TDA model. Table 3 shows the two-way ANOVA summary. Two-way analysis of inconclusive results (Table 4) showed a significant interaction of TDA model and case status, along with a significant main effect for TDA model.

Table 3. Two-way ANOVA summary for correct decisions (TDA model x case status).

Source	SS	df	MS	F	p	F crit .05
Sample	0.024	2	0.000	0.060	.942	3.026
Status	1.794	1	0.012	2.988	.085	3.873
Interaction	0.011	1	0.011	2.652	.104	3.873
Error	1.177	294	0.004		.000	
Total	1.829	298			.000	

Table 4. Two-way ANOVA summary for inconclusive results (model x case status).

Source	SS	df	MS	F	p	F crit .05
Sample	1.350	1	0.014	4.534	.034	3.873
Status	0.626	1	0.004	1.400	.238	3.873
Interaction	0.093	1	0.093	31.231	.000	3.873
Error	0.876	294	0.003		.000	
Total	2.069	297			.000	

A series of one-way post-hoc ANOVAs showed that the three TDA models did not differ significantly for decision accuracy, excluding inconclusive results, for deceptive cases [$F(2,147) = 0.057$, ($p = .945$)], or for truthful cases [$F(2,147) = 0.042$, ($p = .959$)]. Differences in inconclusive rates were not significant for truthful cases [$F(2,147) = 2.112$, ($p = .125$)], but were significant for deceptive cases [$F(2,147) = 4.263$, ($p = .016$)]. The ESS produced the lowest inconclusive rate for deceptive cases.

Decision accuracy and inconclusive rates for the combined deceptive and truthful cases are shown in Figure 2. Unweighted average accuracy for deceptive and truthful cases provides an accuracy estimation that is robust against differences in inconclusive rates, sensitivity and specificity rates, base-rate, or sample size differences between the deceptive and truthful cases. Unweighted accuracy is therefore a numerical index that is easily compared to the accuracy indices from other studies. The two-way ANOVA summary is shown in Table 5.

Figure 2. Unweighted accuracy and unweighted inconclusive rates.

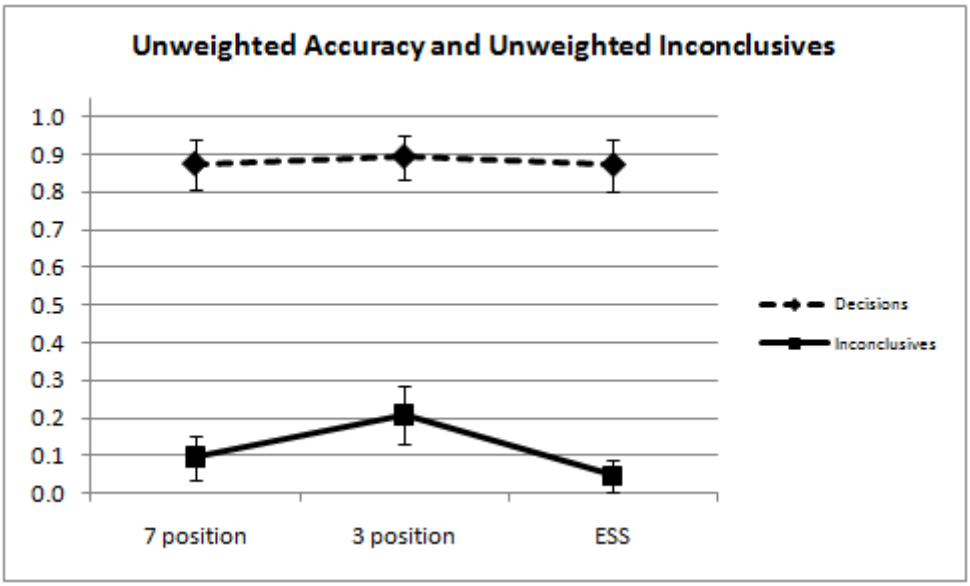


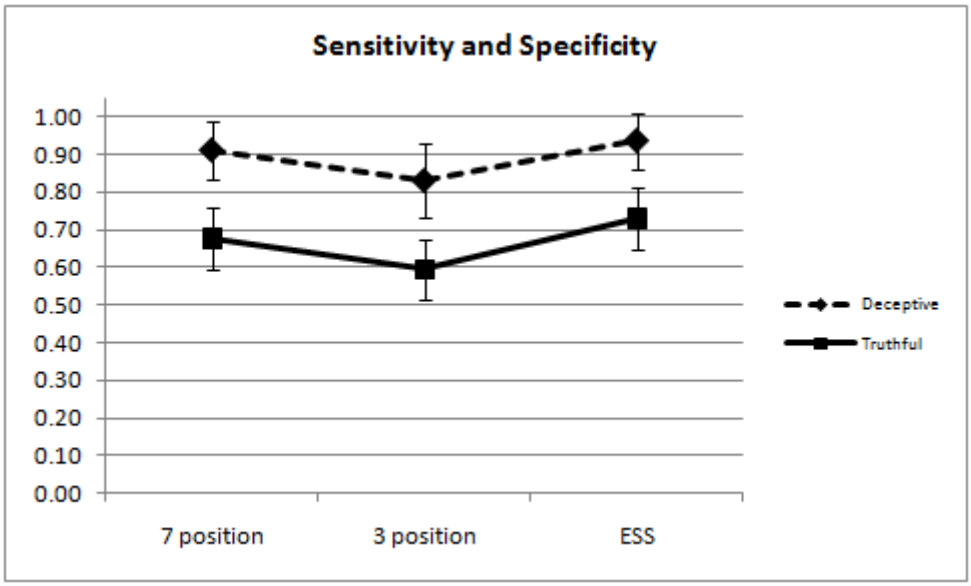
Table 5. Two-way ANOVA summary for TDA model by criterion dimension.

Source	SS	df	MS	F	p	F crit .05
TDA Model	0.445	2	0.004	4.343	.038	3.873
Criterion Dimension	43.586	1	0.291	283.812	.000	3.873
Interaction	0.251	1	0.251	244.880	.000	3.873
Error	0.301	294	0.001		.000	
Total	44.282	297			.000	

The two-way interaction was significant, indicating that the three TDA models can be expected to produce different patterns of correct and inconclusive results. One-way post-hoc ANOVAs showed that the differences in correct decisions were not significant for the three TDA models [F (2,147) = 0.130, (p = .878)]. However, the three TDA models did produce significantly different rates of inconclusive results [F (2,147) = 7.139, (p = .001)]. As described earlier, this difference was significant only for the deceptive cases.

To further explore the relationship between inconclusive results and decision accuracy, an additional two-way analysis (TDA model x case status) was completed for decision accuracy with inconclusives, (i.e., test sensitivity to deception and test specificity to truth-telling). Figure 3 shows the means and 95% confidence intervals for sensitivity and specificity with the seven-position, three-position, and ESS TDA models.

Figure 3. Sensitivity and specificity of DLST examination results with three TDA models.



The interaction of case status and TDA model was significant [F (1,294) = 7.793, (p = .006)], which precluded interpretation of the significant main effects until further analysis. One-way post-hoc ANOVAs showed that differences in test sensitivity to deception were not significant [F (2,147) = 1.709, (p = .185)], while differences in test specificity to truth-telling were approaching a significant level [F (2,147) = 2.752, (p = .067)]. Pairwise contrasts showed that the difference in specificity was not significant for the seven-position and ESS models [F (1,98) = 0.836, (p = .363)] or for the seven-position and three-position models [F (1,98) = 2.000, (p = .160)], but was significant for the three-position and ESS models [F (1,98) = 5.421, (p = .022)].

Discussion

This project was a Monte Carlo simulation designed to study differences in criterion accuracy of seven-position, three-position, and ESS scores of DLST exams that consist of three presentations of two RQs for which the criterion variance is independent, while excluding the pneumograph data from numerical scores.[4] The three TDA models did not differ in overall unweighted accuracy for the simulated DLST exams. However, differences were observed in the inconclusive rates produced by the three TDA models. The ESS produced fewer inconclusive results than the other models, and the three-position model produced more inconclusive results with deceptive cases. Although results suggest that component weighting, as accomplished by the seven-position and ESS TDA models, may be more effective at extracting diagnostic information than the unweighted three-position model, there were no significant differences in test sensitivity to deception, and the simulated DLST exams produced test sensitivity to deception and test specificity to truth-telling that was significantly greater than chance with all three TDA models.

A limitation of this study is that no attempt was made to study DLST criterion accuracy at the level of the individual questions. Previous studies have not supported the hypothesis of highly accurate decisions at the level of the individual RQs. Criterion accuracy of individual RQs should be addressed in future studies.

[4] Whether scored or not during directed-lie exams, pneumograph data is evaluated for indications of behavioral cooperation or non-compliance in actual field examinations.

Another limitation of this study includes the study design as a Monte Carlo study with unknown external validity. Monte Carlo models provide the ability to study a model or problem for which real-world opportunities may be scarce or impracticable. Such is the case with multiple issue PDD screening exams for which the criterion variance is assumed to be independent in the absence of any known incident or allegation. However, Monte Carlo models are only as effective as the design is representative of external conditions, and are always limited by the representativeness of the seed data. Because no suitable seed data exist for directed-lie exams for which the criterion is both known and independent, seed data for this study were adapted from event specific exams by excluding pneumograph scores, which may not be diagnostic with directed-lie exams, and by resampling the seed data to ensure independent criterion variance for the subtotal scores of the simulated exams. No single study can be considered a definitive answer to questions of scientific validity, and these study results should be compared to the results of field and laboratory studies to better understand their generalizability.

Future research should compare DLST screening accuracy with the accuracy of MGQT formats with assumed independent targets. In addition, future research should explore the value of normative data and statistically optimal cutscores for the seven-position and three-position TDA models. Research in TDA methods should continue to evaluate the role of component weighting in the extraction of diagnostic information. Additional research is also needed to better understand the optimal solution for decision rules pertaining to examinations constructed of criterion independent and non-independent examination targets.

With consideration for the obvious limitations of a Monte Carlo study, we suggest that the present result advance our knowledge of test accuracy of DLST exams with seven-position, three-position and ESS scores. These results provide support for DLST exams as capable of providing both test sensitivity to deception and test specificity to truthtelling at rates that are significantly greater than chance, and support the hypothesis that DLST examinations can differentiate deception from truth-telling at rates that are significantly greater than chance when scored with the seven-position, three-position and ESS TDA models. Continued interest in the DLST is recommended, and comparison of these Monte Carlo results with field and laboratory study results is recommended.

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Bell, B. G., Kircher, J. C. & Bernhardt, P.C. (2008). New measures improve the accuracy of the directed-lie test when detecting deception using a mock crime. *Physiology and Behavior*, 94, 331-340.

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 (2006). *Federal Psychophysiological Detection of Deception Examiner Handbook*. Reprinted in *Polygraph*, 40(1), 2-66.

Handler, M., Nelson, R. & Blalock, B. (2008). A focused polygraph technique for PCSOT and law enforcement screening programs. *Polygraph*, 37(2), 100-111.

Handler, M., Nelson, R., Goodson, W. & Hicks, M. (2011). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph*, 39 , 200-215.

Horowitz, S. W., Kircher, J. C., Honts, C. R. & Raskin, D.C. (1997). The role of comparison questions in physiological detection of deception. *Psychophysiology*, 34, 108-115.

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. (2005). Polygraph decision rules for evidentiary and paired testing (Marin protocol) applications. *Polygraph*, 34, 184-192.

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

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

Nelson, R., Blalock, B., Oelrich, M. & Cushman, B. (2011). Reliability of the Empirical Scoring System with expert examiners. *Polygraph*, 40, 131-139.

Nelson, R. & Handler, M. (2010). *Empirical Scoring System: NPC quick reference*. Lafayette Instrument Company. Lafayette, IN.

Nelson, R. & Handler, M. (2012). Monte Carlo Study of Criterion Validity of the Directed Lie Screening Test using the Empirical Scoring System and the Objective Scoring System Version 3. *Polygraph*, 41, 145-155.

Nelson, R., Handler, M. Morgan, C. (2012). Criterion Validity of the Directed Lie Screening Test and the Empirical Scoring System with Inexperienced Examiners and Non-naive Examinees in a Laboratory Setting. *Polygraph*, 41, 176-185.

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

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

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

Research Division Staff (1995a). Psychophysiological detection of deception accuracy rates obtained using the test for espionage and sabotage. DTIC AD Number A330774. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in *Polygraph*, 27, (3), 171-180.

Research Division Staff (1995b). A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope Polygraph and the test for espionage and sabotage question formats. DTIC AD Number A319333. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in *Polygraph*, 26(2), 79-106.

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.