Monte Carlo Study of Criterion Validity of the Directed Lie Screening Test using the Empirical Scoring System and the Objective Scoring System Version 3

Raymond Nelson and Mark Handler

Abstract

Monte Carlo methods were used to calculate normative data and criterion accuracy rates for PDD screening tests conducted using the DLST format for multi-issue examinations. Decision accuracy was significantly greater than chance for all DLST cases and for the deceptive and truthful groups. There were no significant differences in unweighted decision accuracy for two the models. The OSS-3 model produced significantly fewer inconclusive results with truthful cases compared to the ESS. Dimensional profiles of criterion accuracy, including means, standard deviations, and statistical confidence intervals are shown for test sensitivity and specificity, false-negative and false-positive errors, inconclusive rates for deceptive and truthful cases, positive predictive value, negative predictive value, the proportions of correct decisions for deceptive and truthful cases and unweighted decision accuracy. Normative lookup tables for the DLST scores with the ESS are shown in an appendix. Continued interest in the DLST, the ESS and the OSS-3 is recommended.

Introduction

The Directed Lie Screening Test (DLST) is patterned after the Test for Espionage and Sabotage (TES) format for psychophysiological detection of deception (PDD) tests (Research Division Staff, 1995a; Research Division Staff, 1995b) conducted in the context of routine security screening at the US Department of Defense. When used in other screening contexts, for which the test target issues may differ from the investigation targets of government security screening programs (e.g., pre-employment screening of municipal public safety workers, or monitoring offenders in post-conviction treatment and supervision programs), the TES format has been referred to more broadly as a directed lie screening test (Handler, Nelson & and Blalock, 2008). 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 thoroughly reviewed sequence of target questions, anchored by carefully constructed operational definitions that describe the examinee's possible behavioral involvement in the issue or issues of concern. Also included are comparison questions, intended to evoke a measurable response from a truthful person, along with other procedural questions. The DLST differs from other PDD screening formats in that the DLST is conducted with several presentations of all test stimuli within a single test question sequence. In contrast,

[1] *Independence*, in scientific testing of this type, refers to the idea that the criterion status (deceptive or truthful in the case of polygraph testing) of one issue does not affect the criterion status of other target issues. Independence is assumed in both multi-issue screening contexts, and multi-facet investigative contexts, and requires that the individual target issues be evaluated separately before making a categorical determination about the test result as a whole. It is logically and linguistically possible that some PDD screening targets are non-independent (i.e., the criterion state of the issues can affect one another. 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.

traditional PDD testing formats accomplish several presentations or iterations of the test question sequence by repeating the question sequence three to five times while stopping after each presentation of the sequence. The DLST also includes protocols for reducing the occurrence of inconclusive results, including increased requirements for the proportion of non-artifacted and interpretable data, and the immediate review of test questions and repetition of the test question sequence if the results are inconclusive. Although not unique to the DLST, this PDD format is always administered with directed-lie comparison questions.

Development studies on the DLST are limited to TDA methods based on the seven-position test data analysis (TDA) model that was taught at the Department of Defense during the 1990s (Department of Defense, 2006). In the years following the initial development and validation of the DLST, two important trends have occurred in PDD TDA methods. One trend has been the development of computer algorithms to automate some aspects of the TDA process. The Objective Scoring System, version 3, (OSS-3) (Nelson, Krapohl & Handler, 2008)[2] is a powerful open-source algorithm designed to score all types of PDD examination formats including the DLST. A second trend has been an increased emphasis on the use of evidence-based practices, which has resulted in the deprecation and removal of procedures and concepts that lack scientific support. The reduction of scored physiological reaction features, from 23 features to 12 primary and secondary features described in the *Federal Polygraph Examiners Handbook* (Department of Defense, 2006), is an example of the inevitable simplification of methods that an evidence-based approach will foster. The Empirical Scoring System, (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Krapohl, 2010; Nelson & Blalock, in press; Nelson & Handler, 2010; Nelson & Krapohl, 2011; Nelson, Blalock, Oelrich & Cushman, 2011; Nelson et al., 2011; Nelson, et al., 2008),[3] which further reduced the 12 feature model to the three primary features referred to as Kircher features (Dutton, 2000; Harris, Horner & McQuarrie, 2000; Kircher, Kristjansson, Gardner & Webb, 2005; Krapohl & McManus, 1999; Raskin, Kircher, Honts & Horowitz, 1988), is another example of how evidence-based practice requirements can result in the honing of field practices to their robust essentials.

The present study is intended to extend the validation data on the DLST/TES by investigating criterion accuracy with the ESS and the OSS-3. The hypothesis was that the DLST can detect deception and truthfulness at rates greater than chance when scored using the ESS and the OSS-3.

Method

Monte Carlo methods[4] were used to develop normative parameters that would be used to calculate the level of statistical significance for DLST examination results, when scored with the ESS.

Normative mean and standard deviation parameters for truthful and deceptive groups were calculated from a Monte Carlo space of 100 DLST examinations. ESS total scores for the DLST exams in the Monte Carlo space were simulated by standardizing a random number to the mean of the means and standard deviations that were calculated from a single random selection, with replacement, of subtotal scores for each of the seven participants in the Nelson et al. (2008) study. Mean and standard

[2] None of the developers have any financial or proprietary interest in the OSS-3 algorithm, a free and open-source project cross-platform algorithm available to all PDD manufacturers, field examiners and researchers.

[3] None of the developers have any financial or proprietary interest in the ESS, which is available to all PDD professionals.

[4] Monte Carlo models are computer intensive statistical methods used to investigate complex and intangible problems through the use of mathematical simulations based on an emerging base of available knowledge. These methods were first developed by scientists at the Los Alamos National Laboratory who used the code name "Monte Carlo," referring to the casino, for their use of large-scale randomization models during the Manhattan Project.

deviation scores for the seven participants were recalculated for each of 10,000 iterations of the Monte Carlo space. Normative parameters for the Monte Carlo space therefore describe an asymptotic normal distribution whose mean and standard deviation are the bootstrap[5] means of the mean and standard deviation of the ESS scores for the subtotal scores provided by the seven participants in the Nelson et al. (2008) study. 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[6] for the Monte Carlo space.

The Monte Carlo mean for deceptive DLST ESS *subtotal* scores was -2.442 (SD = 3.531), and the Monte Carlo mean for DLST truthful ESS subtotal scores was 2.086 (SD = 3.460). Monte Carlo norms were also calculated for DLST total scores. Because field PDD examinations are scored in integers, not real numbers, normative parameters were truncated to integers before calculating the level of significance for each case in the Monte Carlo model. Appendix A shows the normative lookup data for DLST subtotal scores.

A second Monte Carlo model was then used to calculate the criterion accuracy profile for DLST examinations. The second Monte Carlo space consisted of 100 simulated DLST examinations for which the criterion states for the two target questions in the DLST simulation were set independently by comparing random numbers to a fixed base rate of .293, which was selected as the Šidák correction[7] of the desired base rate of .5 using two independent criteria for each case. The combined base rate was .5 after exhaustive random iterations. Each case was set to a truthful criterion status if neither of the two target questions was randomly set to a deceptive status. Cases were set to a deceptive criterion state if either of the targets were set to a deceptive status. DLST scores for the second Monte Carlo were simulated by standardizing a random number to the deceptive or truthful normative parameters.

A third Monte Carlo model was developed to study DLST criterion accuracy when scored via the OSS-3. The Monte Carlo model was seeded by bootstrapping the subtotal scores in two dimensions, question x case, from the OSS-3 subtotal scores of the confirmed case sample (N=60) from the Nelson and Krapohl (2010) study, excluding the pneumograph data.

Decision alpha[8] was set at a = .1 for truthful classifications and a = .05 for deceptive classifications with the ESS. To maximize test sensitivity, Bonferonni correction was not applied to the alpha cutscore for deceptive classifications using the independent target questions. However, the inverse of the Šidák correction, for

[5] Bootstrap methods entail the building up of distributions based on random and repeated sampling from available sample data so that difficult statistics can be calculated such as confidence intervals to be used for the calculation of estimates of population norms. These population estimates provide a basis for calculating accuracy and other factors. Bootstrapping is computationally intensive, involving thousands of resamplings, and has only become practical with the development of computerized tools. For more information on this approach, see http://www.math.ntu.edu.tw/~hchen/teaching/LargeSample/notes/notebootstrap.pdf.

[6] Seed parameters are those values used to program the Monte Carlo simulation model.

[7] Šidák correction is a mathematical correction for our assumption that individual test questions are independent.

[8] Alpha is a statistical term used to designate a tolerance for a certain proportion of error. Because there is no such thing as a perfect test, some errors are always expected. Alpha is an expression of the target level under which errors should be constrained. An observed rate of error over this level would be considered excessive. Alpha is therefore a decision cutscore, expressed in terms of a probability value. Polygraph examiners traditionally refer to numerical cutscores instead of alpha. Numerical cutscores can be mapped to their corresponding alpha boundaries by using the normative data for deceptive and truthful sample scores for a PDD test question sequence as scored with a specified model for test data analysis.

independent[9] targets, was applied to the alpha for truthful classifications. This was necessary to prevent an increase in inconclusive results, and corresponding deflation of alpha, when calculating the probability that an examination would result in a truthful score in response to all examination targets, while the criterion status of at least one of the investigation targets was actually deceptive. These alpha boundaries corresponded to subtotal cutscores of -3 or lesser for any target questions, resulting in a deceptive classification, and +1 or greater for all target questions, required for a truthful classification.

Because the DLST Monte Carlo target questions were independent, all decisions were made using the spot-score-rule (SSR) (Light, 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, using the SSR, were made only if all subtotal cutscores equaled or exceeded the subtotal cutscores corresponding to the Sidak corrected alpha for truthful decisions. Decision rules for the results of the OSS-3 DLST Monte Carlo model were the same as those for the DLST ESS Monte Carlo model. Both Monte Carlo models were designed to repeat an examination using the same criterion status in the event of an inconclusive result. The ESS and the OSS-3 Monte Carlo models were run for 10,000 iterations.

Results

All statistical analyses were completed with a level of significance set at alpha = .05. Dimensional profiles were calculated for DLST results for the ESS and OSS-3 models, including mean, standard deviation, 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. Unweighted average accuracy for deceptive and truthful cases is an accuracy estimation that is robust against differences in inconclusive rates, sensitivity and specificity rates, and 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. Table 1 shows that the DLST examinations differentiated criterion deceptive from criterion truthful cases at rates that were statistically significantly greater than chance ($p < .05$) for both the ESS and the OSS-3 TDA models, with low overall inconclusive rates for both models.

A series of two-way ANOVAs, TDA model x case status, showed there was a significant main effect for case status for correct decision ($F [1,196] = 12.836$, $p < .001$). Decision accuracy was higher for deceptive cases. In addition, the interaction of TDA model and case status for errors was also statistically significant ($F [1,196] = 8.608$, $p = .004$), along with a significant main effect for case status ($F [1,196] = 12.087$, $p = .001$). False-positive errors occurred more frequently than false-negative errors. However, a series of one-way post-hoc ANOVAs showed that the difference in inconclusives was not significant for the two TDA models for the separate or combined deceptive and truthful cases. Figures 1, 2, and 3 show the mean interaction plots for correct decisions, errors and inconclusive results.

[9] Independence, in scientific testing, refers to assumptions about the degree to which the criterion status of individual target questions influences, and is influenced by, the criterion status of the other target questions. In other words, independence is the degree to which the external factors (i.e., examinee behaviors) that cause the status of responses individual questions to be untruthful or truthful will also affect the untruthful or truthful criterion status of other questions.

Table 1. Mean, (standard deviations) and {95% confidence intervals} for DLST results using the ESS and OSS-3.

	ESS	OSS-3
Sensitivity	.917 (.041) {.836 to .998}	.963 (.026) {.911 to .999}
Specificity	.587 (.071) {.448 to .726}	.683 (.064) {.557 to .808}
FN	.036 (.027) {.001 to .088}	.035 (.025) {.001 to .083}
FP	.253 (.062) {.131 to .375}	.292 (.064) {.167 to .416}
Deceptive Inconclusive	.047 (.032) {.001 to .109}	.003 (.008) {.001 to .018}
Truthful Inconclusive	.160 (.053) {.055 to .265}	.026 (.023) {.001to .070}
PPV	.78 (.054) {.674 to .886}	.767 (.052) {.666 to .868}
NPV	.944 (.041) {.863 to .999}	.952 (.034) {.885 to .999}
D Correct	.962 (.028) {.907 to .999}	.965 (.025) {.916 to .999}
T Correct	.699 (.072) {.558 to .841}	.701 (.065) {.574 to .828}
Unweighted Average Accuracy	.831 (.039) {.755 to .907}	.833 (.035) {.765 to .901}
Unweighted Average Inconclusives	.103 (.030) {.045 to .162}	.014 (.012) {.001 to .037}

Figure 1. Mean plots for correct decisions for ESS and OSS-3.

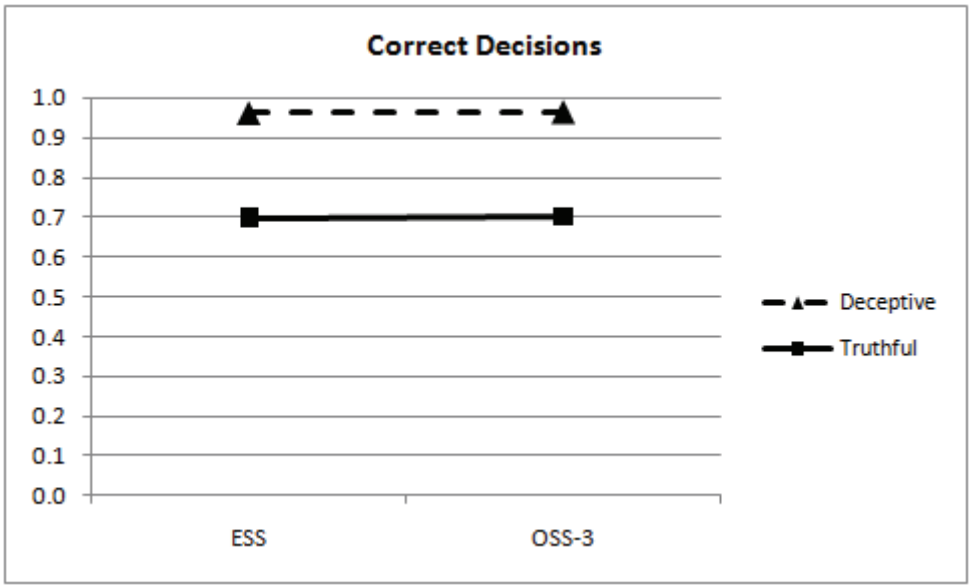


Figure 2. Mean plots for decision errors for ESS and OSS-3.

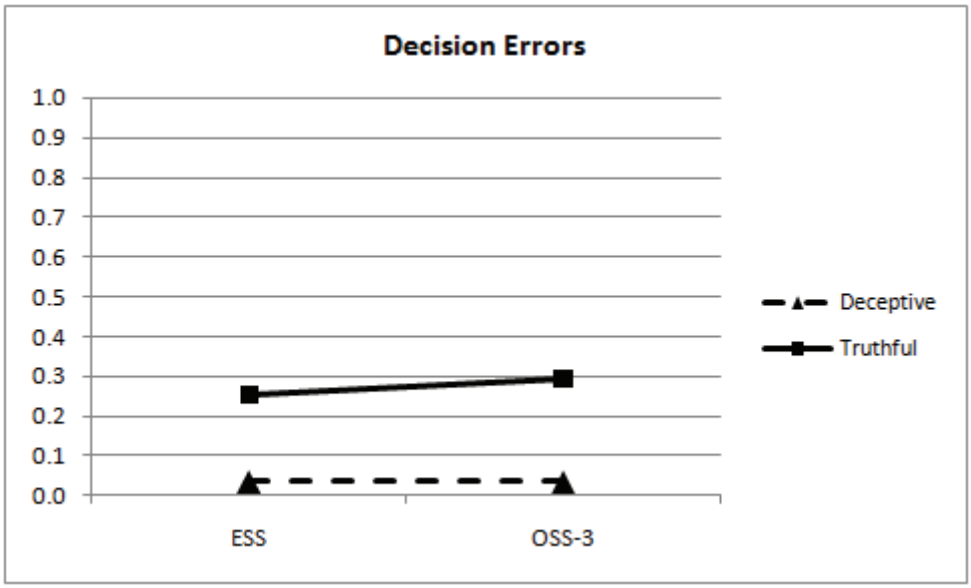
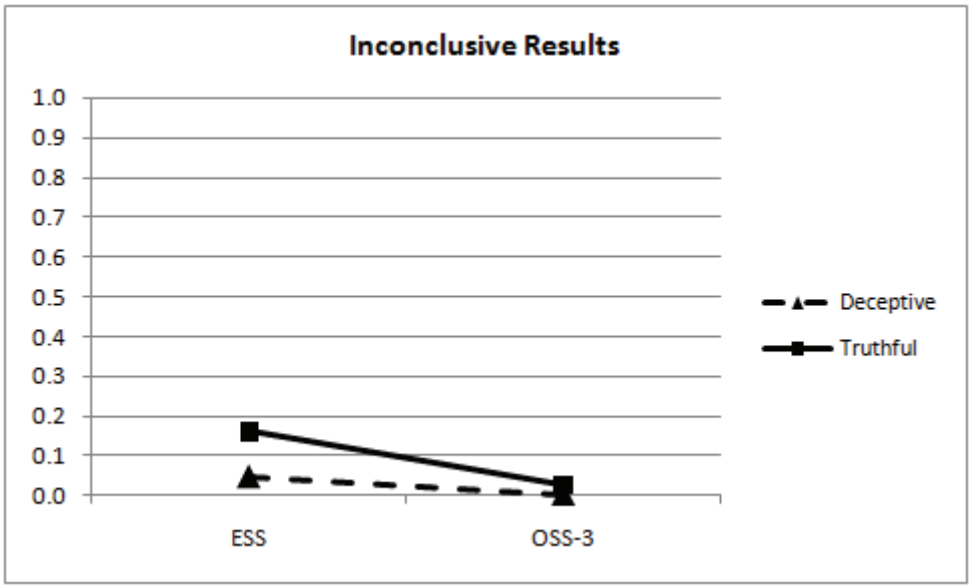


Figure 3. Mean plots for inconclusive results for ESS and OSS-3.



There was a significant interaction of TDA model and case status for inconclusive results (F [1,196] = 91.505, $p < .001$). The OSS-3 algorithm produced significantly fewer inconclusives, and the difference was loaded on the truthful cases. The main effect for TDA model was nearing a significant level (F [1,196] = 3.579, $p = .060$) for inconclusive results. A series of one-way post-hoc ANOVAs showed that the difference in inconclusive results was significant only for the truthful cases (F [1,98] = 5.379, $p = .022$).

Discussion

Validation of a PDD technique, for field diagnostic and field screening purposes, will necessarily involve a combination of two

structural components. The first will be a test question sequence that conforms to valid principles for target selection, test question construction and in-test presentation of the test stimulus. The second structural component of validation involves the TDA method. Either of these components, if ineffective, can greatly affect the overall performance of a PDD technique.

Screening tests present the vexing problem of being conducted, by definition, under circumstances in which there is no known allegation or incident. It is therefore regarded as very difficult or impossible to know the actual state of individual cases in field screening settings. It is likely that research questions pertaining to screening exams will be effectively studied with laboratory studies and with field samples constructed through sub-optimal confirmation methods. To further complicate the challenge, screening exams, in PDD and other settings, are often conducted on multiple simultaneous targets in an attempt to increase the utility of the test. Obtaining research samples of PDD screening exams with extra-polygraphic confirmation of the criterion status of the individual cases has been a challenge that has resulted in little progress towards the validation of PDD screening methods.

Monte Carlo models present an important and necessary solution to the difficulties faced when studying PDD screening methods. The actual criterion state of each case and each question in the Monte Carlo space is known with certainty. Monte Carlo methods attempt to make use of all available related knowledge to study the structural problems and decision models associated with multi-issue screening test performance, in the absence of the ability to study confirmed live data. What is less certain with Monte Carlo methods is the degree to which our prior knowledge is accurate or generalizable to field screening concepts. All studies, including both live studies and Monte Carlo simulations, are limited by the accuracy the *a priori* knowledge-base, and are only as good as the available data.

Test sensitivity (i.e., the ability to identify a high proportion of cases for which the actual criterion state is positive) and a low rate of false-negative errors are the two highest priorities for screening tests. Secondary consideration is given to the overall criterion accuracy of screening tests, along with test specificity (i.e., the ability to accurate identify those cases for which the actual criterion state is negative). Test sensitivity in this study exceeded .900, and the rate of false-negative errors was less than .050. The proportion of correct decisions exceeded .830 for both the ESS and OSS-3 models, with no significant differences in decision accuracy between the two TDA methods. Differences in inconclusive rates were statistically significant for truthful cases ($p < .05$). There were fewer inconclusive results using the OSS-3 algorithm.

No conclusion can be drawn from a single study of any type, and it should go without saying that no single study should ever be considered the final or only answer to the questions about criterion accuracy. As is often the case, additional research is needed. Every study, whether conducted in the field, laboratory, or through Monte Carlo methods, should be evaluated in the context of other studies. It is only through the combined results of multiple studies and through searching for convergent information that we can increase our emerging knowledge base regarding the seemingly intangible issue of criterion accuracy of multi-issue PDD screening exams. It is through the aggregation of results from multiple investigations that researchers and test developers are most likely to identify and develop test construction methods that generalize most effectively to field settings. Results from this study are simply a part of the advancement our current knowledge regarding multi-issue PDD screening exams and the DLST format. With consideration for the acknowledged limitation surrounding all Monte Carlo studies, additional studies using live case data, from both laboratory and field settings, is recommended before making assumptions about the present results as a definitive representation of the criterion accuracy of the DLST format.

Results from this Monte Carlo study support the validity of the hypothesis that the DLST format can differentiate deception and truthfulness at rates significantly greater than chance, when scored with the ESS and OSS-3 TDA models. These results suggest that the

OSS-3 algorithm may be capable of providing important benefits to quality control and training activities involving the review of manual TDA results. Continued interest in the DLST format is recommended, along with continued interest in the ESS and OSS-3 TDA models.

References

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.

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

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. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph*, 39, 200-215.

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.

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

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

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

Nelson, R. & Blalock, B. (In press). Extended analysis of Senter, Waller and Krapohl's USAF MGQT examination data with the Empirical Scoring System and the Objective Scoring System, version 3. *Polygraph*.

Nelson, R., Blalock, B. & Handler, M. (2011). Criterion validity of the Empirical Scoring System and the Objective Scoring System, version 3 with the USAF Modified General Question Technique. *Polygraph*, 40(3), 172-179.

Nelson, R., Blalock, B., Oelrich, M. & Cushman, B. (2011). Reliability of the Empirical Scoring System with expert examiners. *Polygraph*, 40(3), 131-139.

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

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B. & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph*, 40(2), 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(2), 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.

Raskin, D., 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.

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.

Appendix A

Monte Carlo norms for DLST subtotal scores with the Empirical Scoring System

Deceptive Mean = -2.442 (SD = 3.531)
Truthful Mean = 2.086 (SD = 3.460)

Parameters were truncated to integer scores +2 (3) and -2 (3) to produce the following lookup table.

DLST Subtotal Scores			
Truthful Lookup Table (based on the normative distribution of deceptive scores)		Deceptive Lookup Table (based on the normative distribution of deceptive scores)	
Cutscore	Šidák corrected p-value (alpha)	Cutscore	p-value (alpha)
1	.083	-1	.159
2	.047	-2	.091
3	.024	-3	.048
4	.012	-4	.023
5	.005	-5	.010
6	.002	-6	.004
7	.001	-7	.001
8	<.001	-8	<.001