# Criterion Validity of the Directed Lie Screening Test and the Empirical Scoring System with Inexperienced Examiners and Non-naive Examinees in a Laboratory Setting

**Raymond Nelson, Mark Handler and Chip Morgan**

## Abstract

A cohort of inexperienced polygraph examiner trainees from the Iraqi National Information and Investigative Agency (NIIA) and Director General for Intelligence and Security (DGIS) Polygraph Programs used the Directed Lie Screening Test (DLST) with non-naive examinees in a mock espionage scenario as part of their field-training activities. Unweighted decision accuracy was .855 with an inconclusive rate of .086. There were no significant differences in the distribution of deceptive and truthful scores and the distributions of scores from a previous Monte Carlo study on the DLST. A series of two-way analyses showed there were no significant differences between criterion accuracy achieved by ESS scores of DLST examinations and that reported in the development studies on the Test for Espionage and Sabotage.

## Introduction

The directed lie screening test (DLST) (Handler, Nelson & and Blalock, 2008) was developed by the United States Department of Defense as the Test for Espionage and Sabotage (TES) (Department of Defense, 2006; Research Division Staff, 1995a; Research Division Staff, 1995b). Although originally developed for use in federal security screening settings, the TES has been adapted to use in municipal law enforcement selection and post-conviction supervision programs. Prior to the development of this format, PDD screening formats consisted primarily of the family of modified general question techniques (MGQT) which can be traced back to the general question technique (GQT) of Reid (1947). Like all screening tests, the DLST is conducted in the absence of any known incident, known allegation, or known problem. Also like other PDD screening formats, the DLST is designed for use with multiple independent targets for which it is conceivable that an examinee may be involved in one or more target behaviors while remaining uninvolved 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 relevant questions (RQs), comparison questions (CQs), and other procedural questions. Unlike other PDD screening formats, the DLST was designed to maximize testing efficiency with several presentations of all test stimuli within a single test question sequence, without the need to stop or deflate the cuff in between the successive iterations of the test stimuli. Although not unique to the DLST, this examination format is always conducted using directed-lie comparison questions.

Development studies on the 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). Since that time there has been an increased emphasis on evidence-based TDA models and evidence-based practices. This emphasis has led to a reduction of scored physiological features, from 23 or more features to approximately 12 primary and secondary features that have

We are extremely grateful to Sabino Martinez, Pat O' Burke, Akram Sabri Jwad Al NDawi, Mohammed Ahmed Mufeed Kider, Rabea Minhal Araf Al Rubaii, Mohammed Abdul Jabar Al Dulaymi, Mahmood Shaker Raheem, Mohammed Ali Kader, Mina Khadim Al Juburi, Baydaa Hammood Al-Hadeethi, Hassan Falih Hatim (Algaboory), Noor Ismaeel (Al Rubaee), Mohammed Khames Dhari (al Delemi), Asaad Kazim Hassan. Without the commitment of these professionals none of this work would have been accomplished.

been shown repeatedly to be the most robust diagnostic indicators of deception or truth-telling (Harris, Horner & McQuarrie, 2000; Kircher & Raskin, 1988; Kircher, Kristjiansson, Gardner & Webb, 2005; Raskin, Kircher, Honts & Horowitz, 1988). These features have been further reduced to only those considered to be primary features (Dutton, 2000; Harris et al., 2000; Honts & Driscoll, 1987; Kircher et al., 2005; Krapohl & McManus, 1999; Raskin et al., 1988). Nelson, Krapohl and Handler (2008), in a validation study on the Objective Scoring System, version 3, (OSS-3) showed that empirically based manual TDA model, the Empirical Scoring System (ESS), could be used effectively by inexperienced examiners. Other studies have extended the validation data on the ESS (Krapohl, 2010; Nelson, 2011; Nelson & Blalock, in press; Nelson & Krapohl, 2011; Nelson, Blalock & Handler, 2011; Nelson, Blalock, Oelrich & Cushman, 2011; Nelson, Handler, Blalock & Cushman, submitted; Nelson, Handler, Morgan & O'Burke , 2012). Nelson and Handler (2012) used Monte Carlo methods to show that DLST examinations can be interpreted using the ESS with criterion accuracy that is significantly greater than chance.

Previous studies on the DLST included examinee participants who were considered naïve regarding the polygraph test and its administration. Studies have suggested that undetected physical or mental counter-measures can reduce PDD accuracy (Honts, Hodes & Raskin, 1985; Honts, Raskin & Kircher, 1987; Honts, Raskin & Kircher, 1994), but that access to information regarding PDD examinations does not sub-stantially degrade the test accuracy (Honts & Alloway, 2007; Rovner, 1979; Rovner, 1986).

The present study is intended to investigate the criterion validity of ESS results for the DLST when administered to non-naive examinees by inexperienced examiners. The hypothesis was that the DLST can detect deception and truthfulness at rates greater than chance when scored using the ESS.

## Method

Eight polygraph examiner trainees, employed with the Ministry of Defense and Ministry of the Interior in Iraq, participated in this study during their ninth week of training. Three of the participants were female. Ages of the participants ranged from 28 to 42 years. All of the participants had completed four-year college degrees. None of the participants were taking medications for chronic pain, cardiovascular illness, or mental health reasons. Participation in the study was voluntary, and had no effect on the training or employment status of the participants. No harm came to any of the participants as a result of participation in this study.

This study took place in Iraq, in an area known as Forward Operating Base (FOB) Union III. All participants in this study functioned as both PDD examiner and examinee. A laboratory scenario was developed in which study participants were randomly assigned to guilty and innocent groups, with four participants in each group. The principal investigator (RN) was blind to the criterion status of the participants until after the completion of the laboratory and testing activities and data analysis.

Guilty participants were assigned to commit a mock espionage scenario, in which they were told to open an envelope and follow the instructions inside. Instructions required the guilty participants to leave the training room individually at predetermined times and walk to a nearby location at which they were to hand an envelope, marked “secret information” to a man wearing a blue shirt with the number “3” on his sleeve. The man identified himself as a member of an anti-government group. The man wearing the blue shirt was a confederate in the study, and a linguist contractor working in support of US forces and the Iraq government. The envelope marked “secret information” contained a blank business card, and no secret information was actually released to persons associated with anti-government groups as a result of this study. In exchange for the envelope the confederate gave each guilty participant a token that could be exchanged for merchandise at the post exchange (PX). Innocent participants were provided identical envelopes which contained information instructing them to leave the training room individually at predetermined times, walk to a nearby location and then return to the training room. Innocent participants were instructed to answer that they were taking a

break for some exercise if questioned by anyone regarding their presence outside the training room.

Following the completion of the scenario each participant was tested by each of the other participants, using the DLST format. Examination questions, including investigation target questions, directed lie comparison questions, and procedural questions were standardized for all participants. All examinations were conducted in Arabic. Examination targets pertained to providing secret information to persons belonging to anti-government groups, and having unauthorized contact with persons belonging to anti-government groups. Testing activities took place over two days. Examinations were conducted without the use of an acquaintance test. Participants were required to repeat examinations that resulted in inconclusive results. Nine inconclusive examinations were repeated. Four of those examinations resulted in a deceptive classifications after retesting. No post-test discussion was completed following any of the examinations. However, the participants were provided an opportunity to debrief the experience individually and as a group following the completion of all study activities. Participants were required to maintain secrecy regarding their role involvement during study, and there were no discovered lapses or breaches of information for the roles of the participants.

Study participants were given one day of instruction and practice using the DLST before beginning the study activities. The original design was for each of the eight study participants to conduct seven examinations on the other participants, for a total of 56 examinations. However, one participant became sick during the study. This participant was tested by the other participants but was not able to function adequately to participate effectively as an examiner. The participant was released from the study and the remainder of the field PDD training requirements due to the illness. Forty-nine examinations were completed, including 24 examinations of guilty participants and 25 examinations of innocent participants.

DLST examination data were scored using an automated version of the ESS TDA model. The automated ESS consisted of automated measurement of physiological features, automated transformation to integer scores, and automated execution of decision rules. The automated ESS model adhered to the same procedures used when manually scoring DLST PDD examination data, in that each RQ was compared to the stronger of the nearby comparison questions. Because previous studies have suggested that pneumograph data may not be diagnostic with directed lie exams (Bell, Kircher, Bernhardt, 2008; Kircher et al., 2008; Kircher, Packard, Bell & Bernhardt, 2001), pneumograph scores were not included in the automated ESS model for DLST exams. Nelson and Handler (2012) described the development of Monte Carlo norms for DLST examinations scored with the ESS. Appendix A showed the normative data. There were no differences in the frequencies of correct, inconclusicve and erroneous results when the examinations were scored with and without the pneumograph data. Nelson, Handler, Blalock and Cushman (submitted) showed that an automated ESS model can replicate manual ESS scores with no significant differences in criterion accuracy. Alpha was set at .05 for deceptive classifications and alpha = .1 for truthful classifications. The decision rule for the automated ESS model was the spot-score-rule (SSR) (Light, 1999; Swinford, 1999).

ESS cutscores corresponding to these alpha levels were -3 and +1, meaning that 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 ($p < .1$). Bonferonni correction to the alpha cutscore for deceptive classifications is not used with PDD examinations in which it is assumed the investigate target questions are independent. However, an inverse of the Šidák correction for independent issues is 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.

Means, standard deviations, and statistical confidence intervals were calculated for a dimensional profile of criterion accuracy, including: sensitivity, specificity, inclusive results for deceptive and truthful cases, false-positive and false-negative errors, positive predictive value, negative predictive value, percent of correct decisions for the deceptive and truthful cases, and the unweighted means of the percent correct and inconclusive results for deceptive and truthful cases. Distributions of scores were compared to scores obtained from another study (Nelson & Handler, (2012) using multivariate AVOVAs, and the dimensional profile of criterion accuracy was compared, using unbalanced multivariate ANOVAs, to criterion accuracy as reported in development and validation studies on the TES (Research Division Staff, 1995a; Research Division Staff, 1995b).

## Results

All statistical results were evaluated with a level of significance set at alpha = .05.

The mean deceptive subtotal score was -1.271 (SD = 3.131), and the mean truthful subtotal score was 2.667 (SD = 2.299). Comparison of these values with the distribution parameters from an earlier Monte Carlo study of the DLST (Nelson & Handler, 2012) for which the mean deceptive subtotal was -2.442 (SD = 3.531) and the mean truthful subtotal score was 2.086 (SD = 3.460). An unbalanced two-way ANOVA comparison, sample x status, showed an interaction between the sample and case status (F [1,131] = 7.201, p = .008). Unbalanced ANOVA, using the harmonic mean, was necessary due to differences in sample size. The interaction of means can be seen in Figure 1. Post-hoc one-way analysis showed no significant differences between the Monte Carlo distribution and the laboratory study distributions for the deceptive cases (F [1,63] = 2.085, p = .154) or the truthful cases (F [1,63] = 0.731, p = .396), and suggested the interaction was due to the differences in sample size. Table 1 shows the DLST criterion accuracy profile for ESS cores.

**Figure 1. Mean deceptive and truthful scores for Monte Carlo (MC) and laboratory (Iraq) samples.**

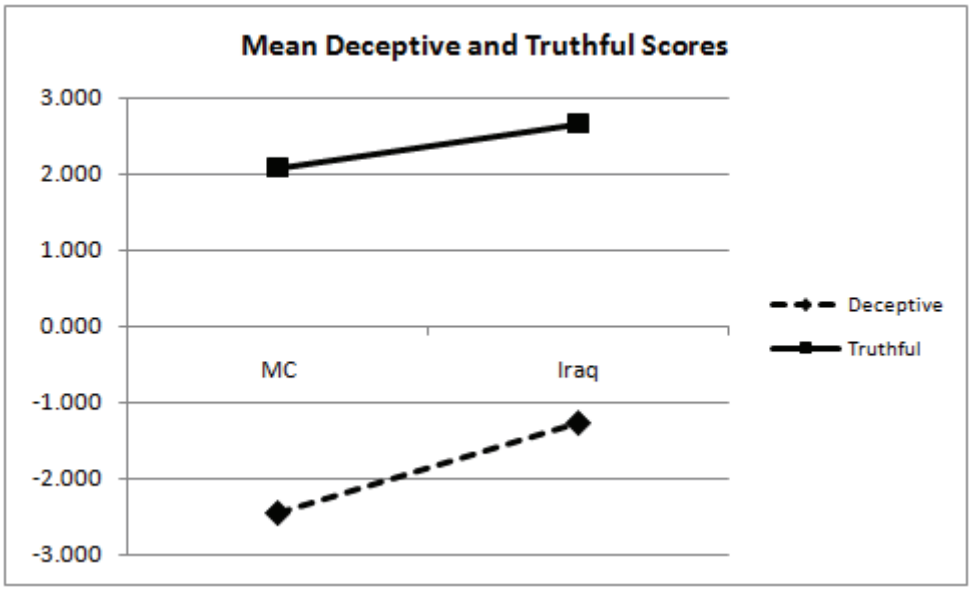


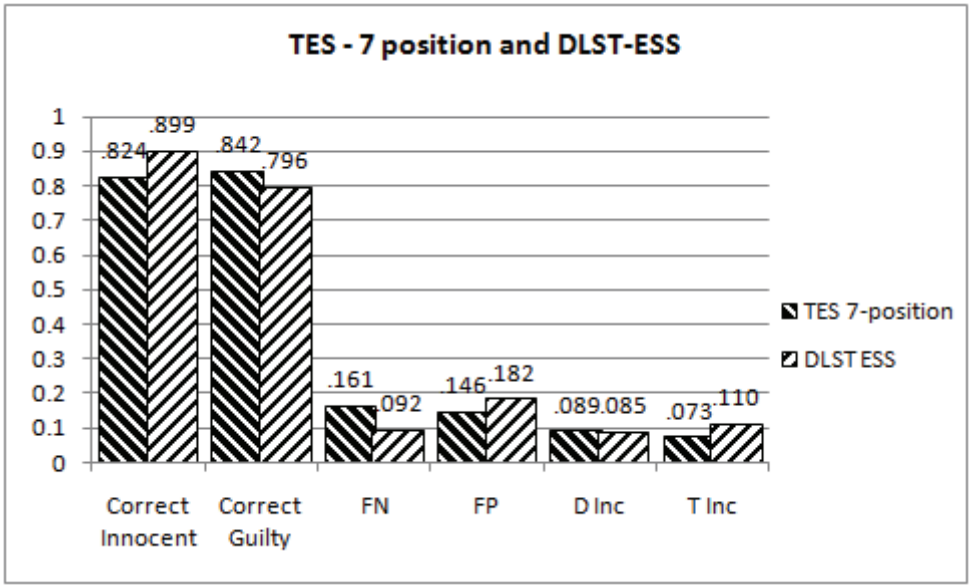
**Table 1. DLST Criterion accuracy**

| | Mean (SD) {95% CI} |
|---|---|
| Unweighted Accuracy | .855 (.036) {.783 to .927} |
| Unweighted Inc | .086 (.026) {.034 to .137} |
| Sensitivity | .628 (.068) {.493 to .763} |
| Specificity | .950 (.030) {.890 to .999} |
| FN Error | .209 (.057) {.096 to .321} |
| FP Error | .039 (.027) {.001 to .092} |
| D Inc | .162 (.051) {.060 to .263} |
| T Inc | .010 (.014) {.001 to .037} |
| PPV | .940 (.041) {.860 to .999} |
| NPV | .819 (.051) {.717 to .921} |
| D Correct | .750 (.067) {.618 to .882} |
| T Correct | .960 (.027) {.906 to .999} |

The proportion of agreement between manual ESS scores and the automated ESS scores was compared to the automated ESS scores using a bootstrap of 1000 iterations. Excluding inconclusive results, the proportion of decision agreement was .911 (SEM = .042) with a 95% confidence range from .828 to .994.

Criterion accuracy results from this study were aggregated together, using weighted averaging, with the results from the earlier Monte Carlo study on the DLST (Nelson & Handler, 2012), and the results were compared to the weighted aggregation of the criterion accuracy profile from the TES development and validation studies (Research Division Staff, 1995a; Research Division Staff, 1995b) in a two-way ANOVA for sample x case status. Figure 2 shows the pattern of mean differences. Results in this analysis were calculated with the inclusion of inconclusive and false-positive error cases that were removed from the reported results of the TES studies. There was a significant interaction of means for correct decisions ($F\ [1,196] = 65.340$, $p < .001$), inconclusives ($F\ [1,196] = 14.137$, $p < .001$) and errors ($F\ [1,196] = 66.034$, $p < .001$). These interactions prevented interpretation of the within group differences without additional analysis.

**Figure 2. Criterion accuracy for aggregated TES development and aggregated DLST ESS studies.**




A series of one-way post hoc ANOVAs showed there were no significant differences in the proportion of correct decisions for deceptive cases ($F$ [1,98] = 0.748, $p$ = .389) or truthful cases ($F$ [1,98 = 0.572, $p$ = .451). Neither were there any significant differences for deceptive cases ($F$ [1,98] = 0.004, $p$ = .952) or truthful cases ($F$ [1,98] = 0.870, $p$ = .353) for the proportions of inconclusive results. Similarly, differences in the proportions of errors was not significant for deceptive cases ($F$ [1,98] = 0.975, $p$ = .326) or truthful cases ($F$ [1,98] = 0.373, $p$ = .542).

## Discussion

These results support the validity of the hypothesis that ESS scores of DLST examinations of non-naive examinees can differentiate deception from truth-telling at rates that are statistically significantly greater than chance ($p$ < .001). There were no significant differences in the distributions of ESS scores of the DLST examinations and the distributions of scores from an earlier Monte Carlo study of ESS scores of DLST exams. In addition, comparison of the aggregated DLST criterion accuracy profile to that of the aggregated TES development and validation studies revealed no statistically significant differences in the criterion accuracy, and suggest the DLST is capable of maintaining greater than chance criterion accuracy levels with non-naive examinees. These results replicate the results of earlier studies which showed that information regarding the PDD examination does not substantially degrade accuracy. Taken together with the results of previous studies involving experienced examiners, these results also indicate that the DLST is a robust technique that can be used effectively by examiners with a wide range of experience.

Although there were no significant differences between the results of this study and those of other studies on the DLST, the degree to which study results will generalize to real world settings is always unknown. Some important differences exist between a laboratory study of the present design and field settings. First, the examinees in this study did not report, and were not observed, attempting to defeat the test results with countermeasure strategies. It can be assumed that deceptive examinees in information security contexts may be more motivated to attempt a variety of measures to alter or disrupt the test result. It can also be assumed that persons involved in espionage or information security breaches are aware of their vulnerability to potential consequences

of great magnitude if they are caught engaging in espionage or caught attempting to falsify their PDD examination results. It is impossible to know with certainty how these forces would manifest themselves in test results in field settings. Although the literature at this time suggests that countermeasure attempts are not highly effective, future studies should continue to investigate the issues of motivation and level of sophistication regarding attempts to defeat or alter the PDD examination results.

An obvious limitation of this study involves the lack of information regarding question specificity. Published studies do not yet support the notion that the PDD examination can differentiate deception and truthfulness between the individual questions within a single exam. That is, PDD examinations have not been shown to be able to determine with high accuracy that an examinee has lied to one or more individual questions while being truthful to others. The practical implication of this has been that PDD examination results are interpreted at the level of the test as a whole, even though the test result may be determined by evaluating responses to individual questions. Future research should continue to investigate polygraph decision models and statistical decision theory regarding tests for which the examination stimuli are assumed to be independent.

Despite the obvious limitations that are inherent to any small-scale study, two important points are worth noting. First, the examiners in this study can be considered to be the least experienced examiners available to participate in a study of this type. Second, the examinees in this study were decisively non-naive, to the point of administering the same examination on every other participant. Examinees in this study can be assumed to have been fully conversant with the investigation targets, comparison questions, psychological basis of testing, and method of test data analysis. It is hoped that the design of the present study will permit some cautious assumptions to be considered, including that experienced examiners may be able to produce better results than this while testing examinees who are almost certainly less familiar with the details of PDD testing. This should be the focus of continued research. Data at the this time suggest continued interest in the DLST as a viable screening mechanism that can be used effectively by examiners with a wide range of experience and with examinees whom may possess a non-naive level of information regarding operational aspects of the PDD test. These data also suggest continued interest in the ESS as a viable method for TDA of screening examinations.

## References


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

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.

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

Honts, C. R. & Alloway, W.R. (2007). Information does not affect the validity of a comparison question test. *Legal and Criminological Psychology*, 12, 311-320.

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

Honts, C. R., Hodes, R. L. & Raskin, D.C. (1985). Effects of physical countermeasures on the physiological detection of deception. *Journal of Applied Psychology*, 70(1), 177-187.

Honts, C. R., Raskin, D. C. & Kircher, J.C. (1987). Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Psychophysiology*, 1, 241-247.

Honts, C. R., Raskin, D. C. & Kircher, J.C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology*, 79, 252-259.

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. & 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., Kristjiansson, 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. (2011). Monte Carlo study of criterion validity for two-question Zone Comparison Tests with the Empirical Scoring System, seven-position and three-position scoring models. *Polygraph*, 40, 146-156.

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. (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(3), 145-155.

Nelson, R., Handler, M., Blalock, B. & Cushman, B. (Submitted). Blind scoring of confirmed federal You-Phase examinations by experienced and inexperienced examiners: Criterion validity with the Empirical Scoring System and the seven-position model. *Polygraph*.

Nelson, R., Handler, M., Morgan, C., O'Burke, P., (2012). Short Report: Criterion validity of the United States Air Force Modified General Question Technique and Iraqi scorers. *Polygraph,* 41, 18-28.

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.

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law and Criminology*, 37, 542-547.

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.

Rovner, L. I. (1979). The effects of information and practice on the accuracy of physiological detection of deception. Unpublished doctoral dissertation.

Rovner, L. I. (1986). Accuracy of physiological detection of deception for subjects with prior knowledge. *Polygraph*, 15(1), 1-39.

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.

# 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 |