Replication and Extension Study of Directed Lie Screening Tests: Criterion Validity with the Seven and Three Position Models and the Empirical Scoring System

Raymond Nelson, Mark Handler, Benjamin Blalock and Nayeli Hernández

Abstract

Two experienced examiners completed blind scoring tasks on 49 Directed Lie Screening Tests (DLST), also known as the Test for Espionage and Sabotage (TES), conducted by seven inexperienced examiners on 8 non-naive examinees who participated in a mock espionage scenario at a forward operating base in Iraq. Seven-position scores, using the US Federal test data analysis (TDA) model, were transformed to three-position and Empirical Scoring System (ESS scores. Monte Carlo models were used to calculate the distributions of seven-position, three-position and ESS scores, and the results were analyzed using multivariate ANOVAs. Unweighted decision accuracy and inconclusive rates using the seven-position scores did not differ significantly from previous studies of the TES at the U.S. Department of Defense. Criterion accuracy for the seven-position, three-position and ESS TDA models was significantly greater than chance. Only the ESS model produced both test sensitivity to deception and test specificity to truth-telling that were significantly greater than chance. The three-position TDA model produced significantly more inconclusive results that were loaded on deceptive cases. The seven-position and ESS scores were found to extract similarly useful diagnostic information from the raw data. Pairwise decision agreement was significantly greater than chance for all models. Results support the criterion validity of the DLST and suggest continued interest in this technique.

Introduction

The directed lie screening test (DLST) (Handler, Nelson & Blalock, 2008; Nelson & Handler, 2012; Nelson, Handler & Morgan, 2012) was based upon the Test for Espionage and Sabotage (TES) (Department of Defense, 2006; Research Division Staff,

We are extremely grateful to the USF-I ITAM-MoD Intelligence Division and ITAM Police Missions who sponsor training for 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 dedicated professionals none of this work would have been accomplished.

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 publication on various polygraph topics. Unrestricted use of this work is granted to polygraph training programs accredited by the American Polygraph Association, or recognized by the American Association of Police Polygraphists or the National Polygraph Association. For information contact raymond.nelson@gmail.com.

Mark Handler is an experienced police examiner and polygraph researcher who helped develop the Objective Scoring System, version 3 and the Empirical Scoring System. Mr. Handler is the research chairperson for the American Association of Police Polygraphists, and has published numerous articles and studies on many aspects of the polygraph. He can be reached at polygraphmark@gmail.com.

Benjamin Blalock is a former federal examiner and APA certified primary instructor, now in private practice. Mr. Blalock can be reached at ben@polygraphtoday.com.

Nayeli Hernández Pimentel is a polygraph examiner in private practice, a member of the APA, and a licensed psychologist in Mexico. Ms. Hernández can be reached at nayeli.psic@gmail.com.

1995a; 1995b) and has been adapted to screening use in public safety selection and post-conviction supervision programs. Prior to the development of this format, Psycho-physiological Detection of Deception (PDD) screening formats consisted primarily of the family of Modified General Question Techniques (MGQT), General Question Techniques (GQT), and the Relevant/ Irrelevant Technique (R/I), which did not include comparison questions and is scored globally or impressionistically and not numerically. The DLST is conducted in the absence of any known incident, known allegation, or known problem and 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 and is always conducted using directed-lie comparison (DLC) 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). Although results of the TES/DLST studies by the Research Division Staff (1995a, 1995b) have been published, neither the data nor any statistical description of the sampling distribution is available for comparison with other samples. Nelson and Handler (2012) used Monte Carlo methods to show that DLST examinations can be interpreted using the Empirical Scoring System (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Nelson, Blalock, Oelrich & Cushman, 2011; Nelson & Handler, 2010; Nelson & Krapohl, 2011; Nelson et al., 2012) and Objective Scoring System version 3 (Nelson, Krapohl & Handler, 2008) with criterion accuracy that is significantly greater than chance. The present study was designed to replicate the results of the seven-position studies conducted by the Department of Defense, and extend our knowledge of criterion accuracy of DLST with the three-position TDA models. The hypothesis was that blind scored results of confirmed DLST examinations from a laboratory study, including results using the seven-position, three-position, and ESS TDA models, can differentiate deception from truth-telling at rates that are greater than chance.

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.

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 where 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 U.S. forces and the Iraqi 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, that 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. Because the examinees were already familiar with the polygraph technique and instrumentation, all examinations were conducted without the use of an acquaintance test after ensuring the proper adjustment and functioning of the instrument. Participants were required to repeat examinations that resulted in inconclusive results. Nine inconclusive examinations were repeated. Four of those examinations resulted in a deceptive classification 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 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 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.

Blind scores were obtained from two examiners using the seven-position model for TDA (Department of Defense, 2006), the third and fourth authors (BB and NH). Both blind scorers were trained at schools accredited by the American Polygraph Association (APA). One blind scorer was trained at the U.S. Department of Defense and is an APA Primary Instructor with five years of experience. The second blind scorer is a native Spanish speaking bilingual examiner, an APA member working in Mexico, and with less than two years of field experience in police polygraph screening programs. Blind scorers worked independently from each other. Seven-position scores were transformed to their corresponding three-position values, and the electrodermal scores were weighted to produce ESS scores.

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 considered a No Significant Response (NSR) result. Any examination with a subtotal of -3 or less or a grand total of -4 or less would be classified as Significant Response (SR). Examinations meeting neither of those conditions would be classified as Inconclusive (INC).

The decision rule for the automated ESS model was the spot-score-rule (SSR) (Light, 1999; Swinford, 1999). Alpha was set at .05 for deceptive classifications and alpha = .1 for truthful classifications. 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 non-deception ($p < .1$). 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.[1] 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, inconclusive 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 percentage of correct decisions and inconclusive results for deceptive and truthful cases. A three-way ANOVA (study x status x criterion dimension) was calculated to compare the decision accuracy and inconclusive rates to those reported by the U.S. Department of Defense (Research Division Staff, 1995a, 1995b). Post-hoc analyses were completed as necessary. A second three-way ANOVA (TDA model, x status x criterion dimension) and post-hoc analyses was completed to compare the unweighted means of the percentage of correct decisions and inconclusive results of the seven-position, three-position, and ESS scores.

Results

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

Sample distributions

Seven-position scores from the two blind scorers produced a mean deceptive subtotal score of -1.833 (SD = 4.099) and a mean truthful subtotal score of 3.670 (SD = 3.443). Three position scores resulted in a mean deceptive subtotal score of -1.458 (SD = 2.784) and a mean truthful subtotal score of 2.470 (SD = 1.853). ESS scores produced a mean deceptive subtotal score of -1.781 (SD = 4.437), and a mean truthful subtotal of 3.636 (SD = 2.917).

Interrater reliability of numerical scores

The proportion of decision agreement was significantly greater than chance for both seven-position and three-position scores. Seven-position scores resulted in a pairwise proportion of decision agreement of .722 (95% CI = .580 to .864). Three-position scores resulted in a pairwise proportion of decision agreement of .761 (95% CI = .602 to .919). Decision agreement did not differ significantly for the two TDA models. ESS scores produced a pairwise proportion of decision agreement of .796 (95% CI = .652 to .940). Decision agreement did not differ significantly for the three TDA models.

Replication

Decision accuracy and inconclusive rates of the seven-position scores were compared to the results reported for seven-position scores in studies reported by the U.S. Department of Defense (Research Division Staff, 1995a; 1995b). Table 1 shows the unweighted average accuracy and unweighted inconclusive rates from the replication and U.S. Department of Defense studies.

Figure 1 shows the mean plots and 95% confidence intervals for the proportions of correct decisions excluding inconclusive results, and inconclusive results of the deceptive and truthful seven-position scores from the present replication and U.S. Department of Defense studies. Table 2 shows the results of a three-way ANOVA (study x

[1] It is often the case that the behavioral details of the investigation target questions are not completely independent.

criterion status x criterion dimension). The three-way interaction between study, status, and criterion dimension was significant, in addition to the two-way interaction between status and criterion dimension, and the main effect for criterion dimension. The main effect comparing the studies was not significant in the three-way analysis, nor was the interaction of study and criterion status.

Table 1. Means, (standard errors), and {statistical confidence intervals} for DLST exams

	7 Position Replication	7 Position DoD Studies
D Correct	0.681 (0.089) {0.507 to 0.856}	0.821 (0.082) {0.66 to 0.982}
T Correct	0.978 (0.021) {0.937 to 1.02}	0.845 (0.082) {0.685 to 1.006}
D Inc	0.145 (0.048) {0.05 to 0.24}	0.087 (0.056) {-0.022 to 0.198}
T Inc	0.039 (0.026) {-0.013 to 0.092}	0.073 (0.051) {-0.028 to 0.175}

Figure 1. Means and 95% confidence intervals for correct decisions and inconclusive results from seven-position replication and US Department of Defense studies

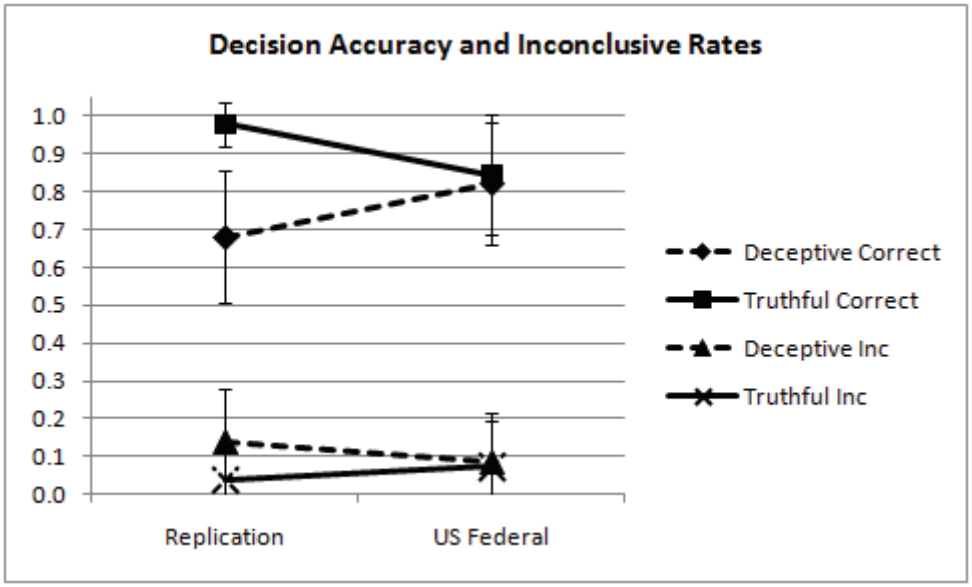

Table 2. Three-way ANOVA summary for Replication and U.S. Department of Defense studies

Source	SS	df	MS	F	p	F crit .05
Criterion dimension	57.761	1	57.761	12708.527	0.000	3.862
Status	0.088	1	0.088	19.297	0.000	3.862
Study	0.001	1	0.001	0.124	0.725	3.862
Criterion dimension x Status	1.441	1	1.441	317.144	0.000	3.862
Status x Study	0.162	1	0.162	35.734	0.000	3.862
Criterion dimension x Study	0.004	1	0.004	0.890	0.346	3.862
Criterion dimension x Status x Study	6.817	1	6.817	1499.793	0.000	3.862
Error	2.036	448.000	0.005			
Total	68.310	455				

A series of one-way unbalanced post-hoc ANOVAs was used to investigate differences between the study results. Differences in decision accuracy were not significant for deceptive cases [F (1,65) = 0.008, (p = .276)] or for truthful cases [F (1,81) = 2.320, (p = .132)]. Differences in inconclusive rates were also not significant for deceptive cases [F (1,65) = 0.357, (p = .552)] or for truthful cases [F (1,81) = 0.280, (p = .598)]. Table 3 shows the unweighted mean decision accuracy and unweighted mean inconclusive rates from the replication and development studies.

Under many circumstances the significant interaction between case status and criterion dimension, as shown in Table 2, would limit the evaluation and reporting of accuracy and inconclusive rates to the separated deceptive and truthful groups, as shown in Figure 1. However, because PDD field examiners and program administrators may be interested in a measure of combined test effectiveness, and because none of the one-way ANOVAs was significant, decision accuracy and inconclusive rates for the combined deceptive and truthful cases are shown in Table 3 and Figure 2. The two-way ANOVA summary (criterion dimension x study) is shown in Table 4, and indicates that the interaction of study and criterion dimension was not significant for the combined groups. The main effect for study was not significant, and the significant main effect for criterion dimension is expected.

Figure 2 shows the interaction plot of the unweighted mean decision accuracy and unweighted mean inconclusive rates from the present replication and the U.S. Department of Defense studies.

Table 3. Means, (standard errors) and {95% CI} for unweighted accuracy and inconclusives

	7 Position Replication	7 Position DoD Studies
Unweighted average accuracy	.829 (.051) {.727 to .930}	.833 (.056) {.723 to .944}
Unweighted average inconclusive results	.090 (.040) {.011 to .169}	.080 (.037) {.006 to .154}

Figure 2. Mean and 95% confidence intervals for decision accuracy and inconclusive rates

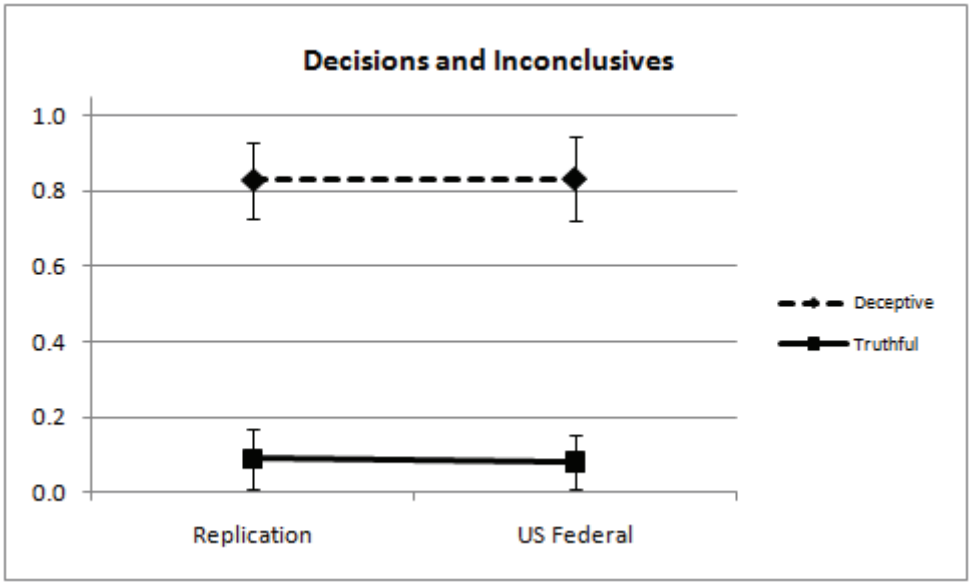

Table 4. Two-way ANOVA summary for decision accuracy and inconclusives of combined deceptive and truthful groups

Source	SS	df	MS	F	p	F crit .05
Study	0.000	1	0.000	0.002	0.964	3.890
Criterion Dimension	27.269	1	0.278	127.847	0.000	3.890
Interaction	0.002	1	0.002	1.103	0.295	3.890
Error	0.418	192	0.002			
Total	27.272	195				

Table 5. Means, (standard errors), and {statistical confidence intervals} for DLST exams

	7 Position Scores	3 Position Scores	ESS Scores
Unweighted Accuracy	.830 (.037) {.756 to .903}	.816 (.043) {.731 to .902}	.858 (.036) {.786 to .929}
Unweighted Inc	.092 (.027) {.038 to .146}	.248 (.042) {.164 to .331}	.123 (.033) {.057 to .188}
Sensitivity	.583 (.070) {.444 to .722}	.415 (.071) {.276 to .555}	.665 (.068) {.532 to .799}
Specificity	.94 (.032) {.876 to .999}	.848 (.050) {.750 to .947}	.839 (.050) {.740 to .938}
FN Error	.271 (.061) {.150 to .392}	.228 (.060) {.110 to .347}	.207 (.058) {.092 to .322}
FP Error	.020 (.020) {.001 to .060}	.009 (.013) {.001 to .036}	.040 (.027) {.001 to .094}
D Inc	.145 (.048) {.050 to .240}	.355 (.069) {.219 to .491}	.126 (.046) {.035 to .217}
T Inc	.039 (.026) {.001 to .092}	.141 (.047) {.047 to .235}	.119 (.045) {.029 to .209}
PPV	.966 (.033) {.900 to 1.032}	.977 (.032) {.913 to 1.041}	.942 (.038) {.866 to .999}
NPV	.775 (.052) {.673 to .878}	.789 (.054) {.683 to .896}	.800 (.057) {.689 to .912}
D Correct	.681 (.089) {.507 to .856}	.644 (.086) {.476 to .813}	.762 (.066) {.631 to .892}
T Correct	.978 (.021) {.937 to .999}	.988 (.016) {.956 to .999}	.954 (.031) {.893 to .999}

Criterion accuracy of seven-position, three-position and ESS scores of DLST exams.

Table 5 shows the dimensional profile of criterion validity for DLST exams when scored with seven position and three position models.

Figure 3 shows the mean plots and 95% confidence intervals for decision accuracy excluding inconclusive results and inconclusive rates for the DLST cases when scored with the seven-position, three-position, and ESS TDA models. Table 6 shows the results of a three-way ANOVA (criterion status x criterion dimension x TDA model) for decision accuracy and inconclusive rates. The three way interaction was significant, along with a significant two-way interaction between criterion dimension and TDA model. Main effects for TDA model, criterion dimension, and status were also significant in the three way analysis. The main effect for criterion dimension is expected as there is no reason why decision accuracy rates should not be different than inconclusive rates.

Figure 3. Means and 95% confidence intervals for correct decisions and inconclusive results

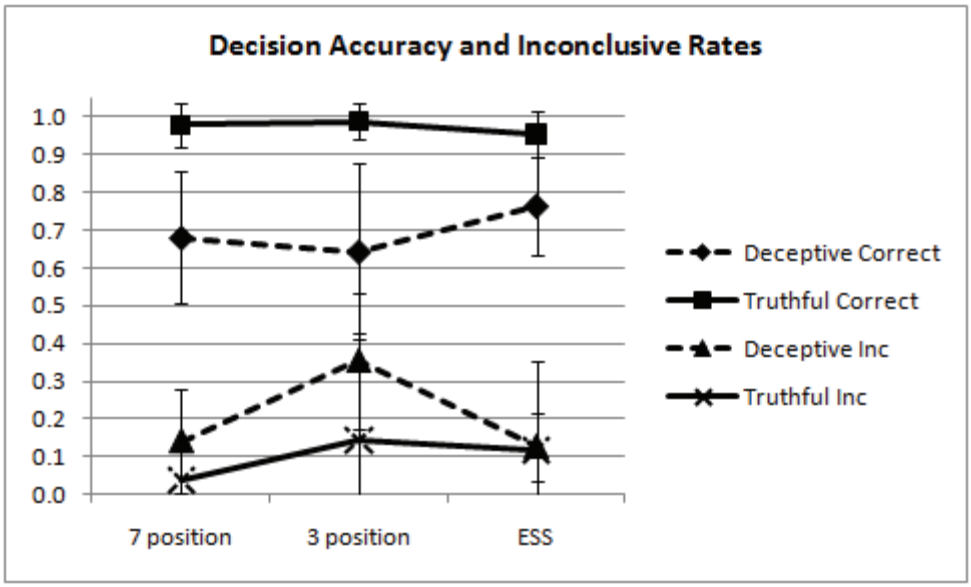

Table 6. Three-way ANOVA summary seven-position three-position and ESS scores

Source	SS	df	MS	F	p	F crit .05
Criterion dimension	33.956	1	33.956	7707.410	0.000	3.875
Status	0.243	1	0.243	55.193	0.000	3.875
TDA Model	0.263	2	0.131	29.847	0.000	3.028
Criterion dimension x Status	3.015	1	3.015	684.452	0.000	3.875
Status x TDA Model	0.013	2	0.007	1.503	0.224	3.028
Criterion dimension x TDA Model	0.481	2	0.240	54.534	0.000	3.028
Criterion dimension x Status x TDA Model	0.402	2	0.201	45.619	0.000	3.028
Error	1.242	282.000	0.004			
Total	39.615	293				

A series of post-hoc one way ANOVAs showed that the differences in decision accuracy, excluding inconclusive results, was not significant for deceptive cases [F (2, 69) = 0.423, (p = .657)] or for truthful cases [F (2, 72) = 0.358, (p = .700)]. Differences in inconclusive rates were not significant for truthful cases [F (2, 72) = 1.025, (p = .367)]. However, differences in inconclusive rates were significant for deceptive cases [F (2, 69) = 3.123, (p = .050)].

Under many circumstances the significant interaction between case status and criterion dimension, as shown in Table 6, would limit the evaluation and reporting of accuracy and inconclusive rates to the separated deceptive and truthful groups, as shown in Figure 3. However, PDD field examiners and program administrators may be interested in a measure of combined test effectiveness. For this reason, and regardless of the significant one-way effect for inconclusive results with deceptive cases, decision accuracy and inconclusive rates for the combined deceptive and truthful cases, are shown in Table 5 and Figure 4. The two-way ANOVA summary (criterion dimension x study) is shown in Table 7.

Figure 4. Mean and 95% confidence intervals for unweighted average decision accuracy and unweighted inconclusives for the seven-position, three-position and ESS TDA models

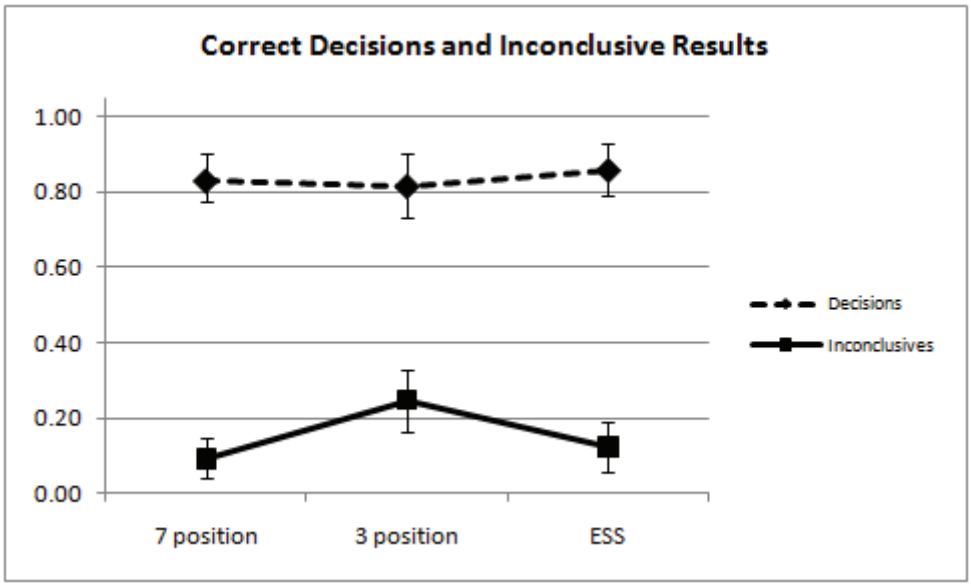

Table 7. Two-way ANOVA Summary

Source	SS	df	MS	F	p	F crit .05
TDA Model	0.249	2	0.003	1.886	0.171	3.874
Criterion Dimension	34.020	1	0.231	171.512	0.000	3.874
Interaction	0.464	1	0.464	343.763	0.000	3.874
Error	0.389	288	0.001			
Total	34.733	291				

The interaction of study and criterion dimension was significant for the combined groups. The significant main effect for criterion dimension is expected and uninteresting. However, main effect for TDA Model could not be interpreted due to the significant interaction. One-way post-hoc ANOVAs showed that the difference in decision accuracy was not significant for the three TDA models [F (2,144) = 0.304, (p = .738)], while the difference in inconclusive rates was significant [F (2,144) = 5.712, (p = .004)]. The three position TDA model produced more inconclusive results than the other models. Pairwise ANOVA contrasts of the TDA models revealed that the difference in inconclusive rates was significant for the three-position and ESS models [F (1,46) = 4.913, (p = .032)]. The three-position model produced significantly more inconclusives. The difference in inconclusive rates between the seven-position and three-position models was approaching a significant level [F (1,46) = 3.359, (p = .073)]. Differences in inconclusive rates for the seven-position and ESS models were not significant [F (1,46) = 0.314, (p = .860)].

Discussion

Results of this study replicate the unweighted DLST decision accuracy and inconclusive rates of DLST examinations, as reported in previous studies by the U.S. Department of Defense. All three TDA models, seven-position, three-position, and ESS, produced criterion accuracy that was significantly greater than chance. Interrater decision agreement, excluding inconclusive results, was significantly greater than chance for all three TDA models. Overall decision accuracy, excluding inconclusive results, did not differ significantly for the seven-position, three-position, or ESS models. However, only the ESS model produced both test sensitivity to deception and test specificity to truth-telling that were significantly greater than chance. The three-position model produced significantly more inconclusive results than the ESS, and the difference in inconclusive results was approaching a significant level for the seven-position and three-position models. Differences in inconclusive rates were significant only for the deceptive cases, suggesting that the component weighting achieved by the seven-position and ESS models may increase test sensitivity, making these models more effective at extracting diagnostic information indicative of deception. The absence of significant differences between the seven-position and ESS models suggests that these models are similarly effective at extracting and using diagnostic information. The inconclusive rates produced by the three-position model may be considered excessive for use in field PDD programs. It is possible that the use of normative data, optimal cutscores and improved decision rules could improve decision accuracy for the three-position TDA model. Additional research is recommended in this area.

Although the sampling distributions of scores are not available for direct comparison, these results suggest that differences may exist between the distributions of scores from this study and those of previous studies (Research Division Staff, 1995a, 1995b). Scores from the present replication were more effective with truthful cases, while scores from the previous studies appear to be more effective with deceptive cases. However, differences in results are not statistically significant, and the exact cause of these observed differences will remain unknown without further study. It is possible that these differences are the result of differences in study design, or to differences in internal and external motivation for the study participants. It is also possible that these differences are the result of differences in language and culture, classroom or professional relationships among the study participants, examiner experience, or the degree of naivety of the examinees regarding the PDD examination. Previous studies on the DLST involved experienced examiners and presumed naïve examinees, while the present replication was conducted under adverse circumstances, with inexperienced examiners testing examinees who were explicitly non-naive.

Despite the acknowledged differences, these results replicate the results of earlier studies, which showed that information regarding the PDD examination does not substantially degrade accuracy. In addition to the presumed adverse condition of testing non-naive examinees, the environmental conditions of the data collection were also adverse. The study location known as FOB Union III is located in a war zone. Indeed the

study and training facility was subject to an explosive rocket attack during the period of the study. The exact effect of these stressors on the performance of the study participants cannot be known.

An interesting aspect of these results, and the fact that examinations were conducted in the Arabic language, is that they demonstrate the ability of a highly standardized PDD test format, including RQs and DLC questions, to transcend language and cultural barriers and remain effective.

Limitations of the present study include the small cohort of scorers, small sample size, and the unknown degree to which the results of laboratory examination data obtained by inexperienced examiners and non-naive examinees are generalizable to field settings. One noteworthy result is that the pattern of inconclusive results in this study is contrary to that of the general trend in the literature. This study resulted in more effective performance with truthful cases and more inconclusive results with deceptive cases. Most previous studies produced the opposite pattern, and the exact reasons for these differences are unknown. Replication of this study and continued research is recommended.

An additional limitation to this study was that no attempt was made to investigate decision accuracy at the level of the individual RQs. Previous research has not supported the hypothesis of highly accurate decisions at the level of the individual questions, and decisions in this study were made at the level of the test as a whole when evaluating subtotal scores for individual questions. Future research should investigate DLST decision accuracy at the level of the individual question subtotals.

Regardless of these differences and limitations, the results of this study differ minimally from those of previous studies on the DLST, and support the validity of the DLST as capable of differentiating deception and truth-telling at rates that are significantly greater than chance under adverse circumstances. Continued interest in the DLST is recommended.

Acknowledgments

Special thanks to Chip Morgan and Sabino Martinez without whose assistance and professionalism this study would not have been possible.

References

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. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph*, 39(4), 200-215.

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.

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., & 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(3), 176-185.

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., Handler, M., Shaw, P., Gougler, M., Blalock, B., Cushman, B., Russel, C., & Oelrich, M. (2011). Using the Empirical Scoring System, *Polygraph*, 40(2), 67-78.

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

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). 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.

Research Division Staff. (1995b). 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.

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.