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

Raymond Nelson, Mark Handler, Ben Blalock and Barry Cushman

Abstract

This paper describes the results of a blind-scoring study of criterion accuracy of two-question ZCT examination. Fifteen scorers in two cohorts completed scoring tasks on a sample of You-Phase exams (N = 22) taken from the Department of Defense confirmed case archive using different scoring models including the seven-position and three-position methods, the ESS, and the OSS-3 computer algorithm. One cohort consisted of scorers with mixed levels of experience, and another consisted of only experienced scorers. High correlations are reported (r = .961) between ESS scores and seven-position scores that are transformed to ESS scores, and two-way ANOVAs showed no significant differences between the distributions of ESS and transformed scores. Additionally, there was no significant difference between the manual ESS scores and previous scores for two relevant question exams using a Monte Carlo model. Criterion accuracy for the ESS, and OSS-3 computer algorithm exceeded 90%. Pairwise decision agreement for ESS results was over 85% for inexperienced scores and exceeded 93% for experienced scorers. Excluding inconclusive results, pairwise decision agreement for the OSS-3, the automated ESS model, and the results of the averaged ESS scores of the 15 study participants was over 98%. Experienced scorers produced significantly fewer false-positive errors than inexperienced scorers. No significant differences were found in decision accuracy with inconclusives, errors or inconclusive results for ESS scores and those from an automated version of the ESS. Continued interest in the You-Phase format, and continued interest in the ESS and seven-position TDA models are recommended.

Introduction

The You-Phase technique is a commonly used single issue test format for psychophysiological detection of deception (PDD) exams. You-Phase examinations consist of two investigation target questions that describe the examinee's behavioral involvement in a single known incident or allegation, along with three comparison questions, in addition to other procedural questions that are not numerically scored. The name of the technique ("You-Phase") is a reference to the basic form of the relevant stimulus question, "Did you do it?" in which both questions describe the examinee's direct involvement in the issue of concern.

The You-Phase format exists today in two closely related versions: the version taught at the National Center for Credibility Assessment (Department of Defense, 2006b) and several polygraph schools accredited by the American Polygraph Association, and the version originally developed by Cleve Backster. Both versions have their origins in the work of (Backster, 1963) and (Reid, 1947). There are no substantive differences in the sequence of test questions, principles for target selection or question formulation for these two versions. The two versions differ in their method of test data analysis, including features, transformation rules, decision rules, and cutscores. The Backster version of the You-Phase technique was used in a series of studies on the effects of countermeasures on PDD examinations (Honts & Hodes, 1982; Honts & Hodes, 1983; Honts, Hodes & Raskin, 1985). Meiron, Krapohl & Ashkenazi (2008) studied the Backster "either-or" rule used with the Backster You-Phase technique. None of these studies were intended to address the issue of criterion accuracy. The You-Phase technique is supported by a complete procedural description (Department of Defense, 2006b), and by favorable opinions anchored in decades of case

experience and anecdotal evidence. Nelson (2012) studied criterion accuracy of the You-Phase technique using a Monte Carlo model.

Criterion validity of field PDD techniques is influenced by a combination of the test question sequence - which should conform to valid principles for target selection, question formulation, and in-test presentation of the question sequence - and the method for test data analysis. Both of these will have a substantial impact on test performance. Other variables may also affect criterion accuracy, including the suitability of the examinee and the effectiveness of the pretest interview; however, these were not the focus of the present study.

The present study was designed to investigate the criterion accuracy of blind-scores of confirmed You-Phase examinations conducted during field investigations. The hypothesis was that blind scored results of confirmed You-Phase exams from field investigations, including results using the seven-position and three-position TDA models (Department of Defense, 2006a; Department of Defense, 2006b), the ESS (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2011; Krapohl, 2010; Nelson & Handler, 2010; Nelson & Krapohl, 2011; Nelson, Blalock, Oelrich & Cushman, 2011a; Nelson et al., 2011b; Nelson, Krapohl & Handler, 2008)[1] , and the Objective Scoring System, version 3[2] (OSS-3) (Nelson et al., 2008) computer algorithm, can differentiate deception from truth-telling at rates that are better than chance.

Method

Data were obtained from two cohorts of scorers, the first being a cohort of 10 scorers with mixed levels of experience consisting of eight inexperienced examiner trainees with the Iraqi military or police, and two experienced examiners, both former certified primary instructors by the American Polygraph Association (APA). These scorers performed blind scoring tasks using the seven-position TDA model (Department of Defense, 2006a). Two-hundred and twenty (220) seven-position examination scores were obtained from the mixed-experience cohort and these scores were transformed to corresponding three position values, and further transformed to ESS scores. A second cohort consisted of experienced scorers, including four experienced Iraq examiners, who have been working in the field for approximately three years and are estimated to have collectively conducted in excess of 2000 field exams, and two experienced US examiners, both formerly certified APA primary instructors.

The second cohort provided only ESS scores for the study. One participant scored the sample cases using the ESS and then rescored the sample cases using the seven-position model approximately two weeks later. This participant was not provided with any feedback or the confirmation status for the sample cases and remained blind to the criterion during both scoring activities. One hundred thirty-two (132) You-Phase examination scores were obtained from the second cohort. One experienced scorer, participated in both cohorts, providing both seven-position and ESS scores, and these scores were not included in calculations which compared the ESS scores from the two cohorts. So in total, fifteen scorers participated in this study: eight inexperienced examiner trainees, and seven experienced examiners. No scorers were provided with the confirmation status for the sample cases and remained blind to the criterion during all scoring activities. All worked independently and without assistance from other participants. Participation in this study was voluntary. There were no incentives or rewards offered to the study participants.

Sample data were a matched random sample (N = 22) of You-Phase examinations selected from the confirmed case archive at the Department of Defense. Eleven confirmed truthful examinations met the selection crite-

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

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 field examiners and researchers.

ria, which involved physically healthy adult criminal suspects, reportedly not taking psychotropic medications, whose examinations consisted of three test charts. Eleven matching confirmed deceptive examinations were randomly selected. Examinee data were completely anonymous, and selection into the study sample had no effect on the criminal investigation or case outcome. Examinations consisted of two relevant questions regarding a single issue, three comparison questions, and other procedural questions as required by the You-Phase technique. The sample cases were confirmed as deceptive via a combination of extra-polygraphic evidence and confession, or extra-polygraphic evidence or confession which inculpated an alternative suspect, thus exonerating the examinee. All examinations were conducted by US Federal and local law enforcement agencies, using the procedures described by the (Department of Defense, 2006b), and were submitted to the Department of Defense for review and inclusion in a confirmed case archive. Examination results from the original examiners were not 100 percent accurate.

Data were also evaluated using the OSS-3 algorithm (Nelson et al., 2008), a free, open-source, and cross-platform statistical algorithm designed to calculate a statistical classifier for a wide variety of PDD examination techniques including the You-Phase format.

Bootstrap Monte Carlo methods and multivariate ANOVAs were used to evaluate the scores and results of the confirmed You-Phase field examinations for the different TDA models.

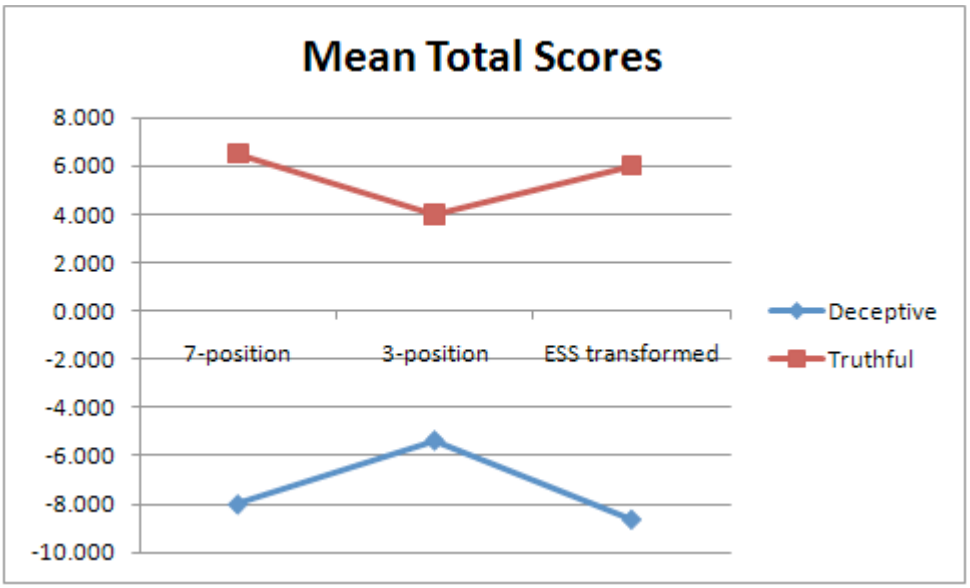
Results

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

Generalizability of the Sampling Distributions. Mean totals were calculated for the 220 numerical scores from the seven-position, three-position and ESS numerical transformations. The mean total deceptive score for the seven-position model was -7.991 (SD = 6.733), and the mean total truthful score was 6.514 (SD = 6.680). Transformation of seven-position scores to their corresponding three-position counterparts resulted in a mean total deceptive score of -5.394 (SD = 4.219) and a mean total truthful score of 3.982 (SD = 4.432). Additional transformation to ESS scores, accomplished by doubling the EDA scores of the transformed three-position scores, resulted in a mean total deceptive score of -8.606 (SD = 5.842), and a mean total truthful score of 6.018 (SD = 7.107).

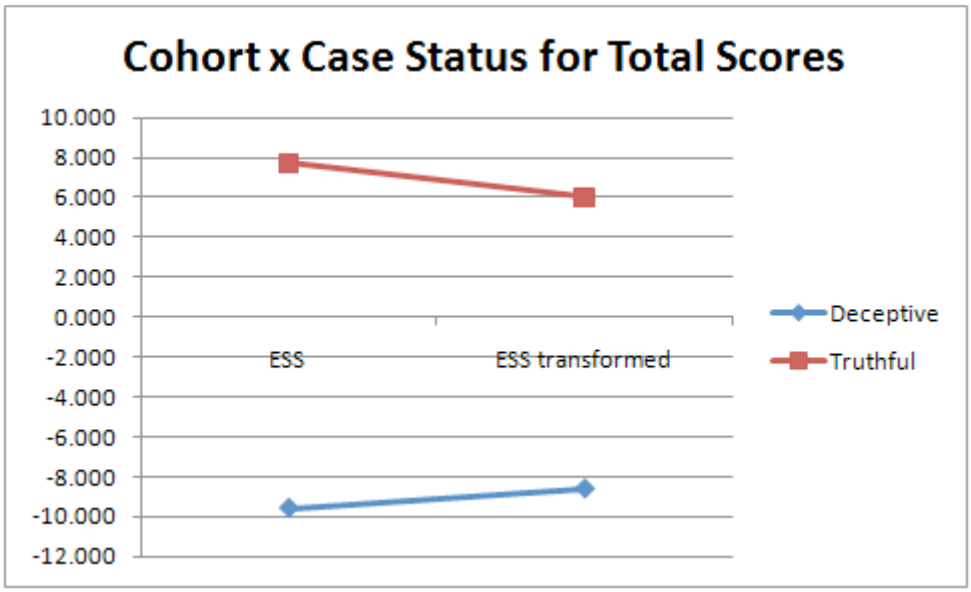
A two-way ANOVA, model x status, showed there was a significant interaction for mean scores ($F_{1,654} = 27.331$, $p < .001$). Figure 1 shows the mean plots of absolute total scores, and illustrates that the three-position transformation model produced total scores of weaker absolute value (i.e., closer to zero). One-way post hoc ANOVAs showed no significant differences within the deceptive or truthful distributions for the three transformation models.

Figure 1. Mean plot of absolute total scores for seven-position, three-position, and ESS models.



Evaluation of the 132 ESS scores obtained from the six experienced examiners who provided ESS scores produced a mean total deceptive score of -9.576 (SD = 4.769) and a mean total truthful score of 7.727 (SD = 6.676). Figure 2 shows that the ESS scores from the experienced cohort were of slightly larger absolute value (i.e., further from zero), compared to the transformed ESS scores from the mixed experience cohort. A two-way ANOVA comparison, cohort x status, of absolute total scores showed no significant interaction and no significant main effects. These results indicate that ESS scores and transformed ESS scores approximate each other reasonably well.

Figure 2. Mean plot of ESS from the six experienced examiners and transformed ESS scores from the experienced and mixed experience cohorts.

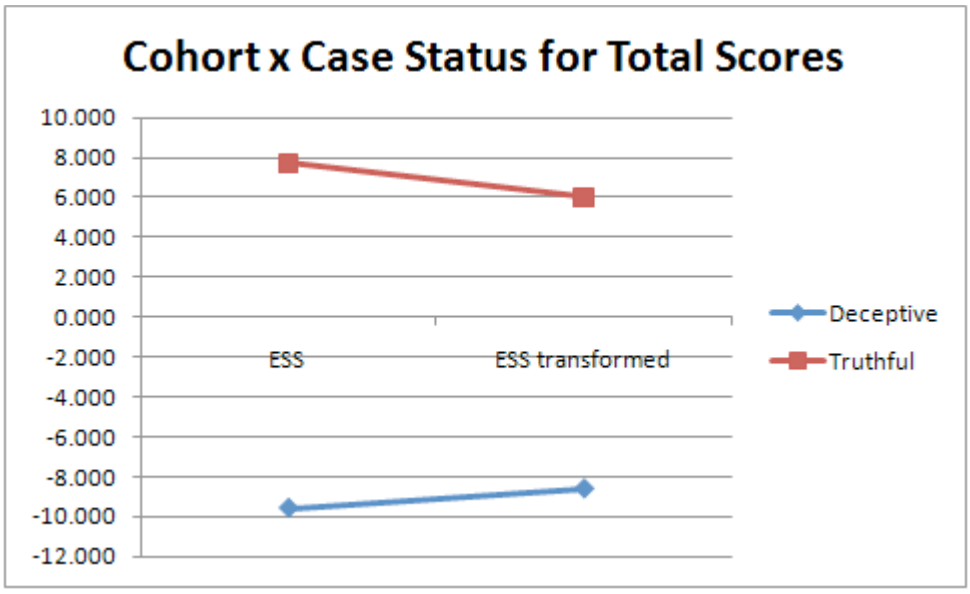


Total scores for nine of the ten scorers in the mixed experience cohort were averaged for each case, and total scores for five of the six scorers in the experienced cohort were also averaged for each case. Scores of one participant were removed from these averages because one experienced examiner scored the data in both cohorts, providing both seven position and ESS scores. Comparison of the mean total ESS and transformed ESS scores resulted in a correlation coefficient of .961.

Evaluation of ESS scores and transformed ESS scores provided by the one experienced scorer who scored the cases using both models, removed from the averages for both cohorts, resulted in a correlation coefficient of r = .968, with perfect agreement among the decisions based on the two sets of scores. A two-way repeated measures ANOVA showed only an expected effect for criterion status ($F_{1,40} = 280.414$, $p < .001$), but no main effect for replication, and no interaction of repetition x status.

It is axiomatic that all sampling distributions are biased and provide an imperfect representation of the population. Generalization of study results to the population is justified only when it is reasonably certain that the degree of sampling bias is quantifiable and tolerable. Because the exact population distribution will always remain unknown and unavailable for comparison with sampling distributions, sampling distributions are commonly evaluated for their representativeness through comparison with other sampling distributions, and through the process of replication. To evaluate the representativeness of the data obtained during this study, the sample distribution of absolute total ESS scores from the experienced cohort was compared using two-way unbalanced ANOVA, sample x case status, accounting for differences in sample size, to the asymptotically normal Monte Carlo distribution described by Nelson (2012), for which the mean total deceptive score was -6.685 (SD = 6.881) and the mean total truthful score was 6.735 (SD = 6.045). Figure 3 shows the mean interaction plot for ESS scores from the experienced cohort and the Monte Carlo normative distribution described by Nelson (2012), and reveals that both truthful and deceptive scores were further from zero for ESS scores from the experienced cohort, compared to the Monte Carlo normative parameters. There were no significant main effects for case status or sample. However, the interaction of case status and sample was approaching a statistically significant level ($F_{1,118} = 2.408$, $p = .123$). These results suggest that the sampling distribution of ESS scores obtained from the experienced cohort is reasonably represented by the Monte Carlo norms.

Figure 3. Mean plot for interaction of ESS scores and Monte Carlo norms.



Criterion Validity. A dimensional profile of criterion accuracy was calculated for the seven-position and three-position TDA models, using traditional federal rules (Department of Defense, 2006a & 2006B) and two-stage decision rules (Krapohl, 2005; Krapohl & Cushman, 2006; Senter, 2003; Senter & Dollins, 2008a; Senter & Dollins, 2008b), including sensitivity, specificity, inconclusive results for deceptive and truthful cases, false-positive and false-negative errors, positive predictive value, negative predictive value, percentages of correct decisions for the deceptive and truthful cases, and unweighted mean of the percentage of correct decisions without inconclusives[3], and unweighted inconclusive rates for the deceptive and truthful cases. Table 1 shows the criterion accuracy profile of the mixed experience cohort of 10 participants who provided 220 scores for the confirmed You-Phase field sample (N = 22) for the seven-position and three-position using both two-stage and traditional TDA models. Three-position scoring produced unweighted decision accuracy that was equivalent to that of the seven-position model, though with a greater proportion of inconclusives. Test specificity for the three-position model was not greater than chance for both traditional decision rules and two-stage rules.

3 Unweighted decision accuracy excludes inconclusive results, and is calculated as the arithmetic mean of TP / (TP + FN) and TN / (TN + FP). This statistic makes no assumptions about base-rates. It is robust against unbalanced cell sizes for the deceptive and truthful sample cases, which could affect the observed proportion of correct, inconclusive and erroneous decisions if the sample results are calculated for the combined deceptive and truthful cases.

Table 1. Mean, standard deviation, and 95% confidence intervals for seven-position and three-position TDA models.

	Seven-position (traditional rules)	Seven-position (two-stage rules)	Three-position (traditional rules)	Three-position (two-stage rules)
Unweighted Accuracy	.874 (.034) {.806 to .942}	.883 (.032) {.820 to .947}	.877 (.038) {.801 to .953}	.878 (.038) {.802 to .954}
Unweighted Inc	.130 (.033) {.063 to .196}	.108 (.031) {.047 to .169}	.235 (.042) {.151 to .319}	.220 (.039) {.142 to .299}
Sensitivity	.845 (.050) {.747 to .943}	.844 (.050) {.746 to .942}	.826 (.053) {.723 to .930}	.827 (.053) {.723 to .932}
Specificity	.677 (.063) {.553 to .802}	.730 (.063) {.606 to .854}	.530 (.070) {.392 to .669}	.555 (.070) {.417 to .693}
FN Error	.037 (.026) {.001 to .090}	.036 (.025) {.001 to .087}	.027 (.024) {.001 to .074}	.028 (.024) {.001 to .076}
FP Error	.179 (.054) {.073 to .285}	.171 (.053) {.067 to .276}	.143 (.050) {.044 to .242}	.147 (.052) {.044 to .249}
D Inc	.116 (.045) {.028 to .205}	.119 (.045) {.029 to .209}	.145 (.049) {.047 to .243}	.143 (.049) {.047 to .24}
T Inc	.143 (.049) {.046 to .240}	.097 (.042) {.014 to .180}	.325 (.068) {.191 to .458}	.297 (.064) {.170 to .424}
PPV	.825 (.053) {.72 to .929}	.830 (.052) {.727 to .933}	.851 (.052) {.749 to .953}	.848 (.052) {.745 to .951}
NPV	.947 (.037) {.873 to .999}	.952 (.032) {.888 to .999}	.951 (.042) {.867 to .999}	.951 (.041) {.87 to .999}
D Correct	.957 (.03) {.897 to .999}	.958 (.029) {.901 to .999}	.967 (.028) {.912 to .999}	.966 (.028) {.911 to .999}
T Correct	.791 (.061) {.671 to .910}	.809 (.058) {.694 to .924}	.787 (.071) {.647 to .926}	.79 (.072) {.649 to .931}

Table 2 shows the accuracy profile for all 15 study participants, mixed and experienced cohorts combined, who provided a total of 330 ESS scores for the confirmed You-Phase field sample (N = 22), along with the accuracy profile for an automated version of the ESS and the OSS-3 computer algorithm. For this sample the OSS-3 algorithm produced an overall level of decision accuracy that was not different than the other models, though with a very high level of sensitivity to deception (.980), almost zero inconclusives (.009), and a higher false-positive error rate (.181) than the other models.

Table 2. Mean, standard deviation, and 95% confidence intervals for ESS, automated ESS, and OSS-3 criterion accuracy.

	ESS	Automated ESS	OSS-3
Unweighted Accuracy	.901 (.030) {.841 to .962}	.944 (.023) {.897 to .990}	.903 (.029) {.846 to .961}
Unweighted Inc	.096 (.028) {.040 to .152}	.090 (.028) {.033 to .147}	.009 (.009) {.001 to .029}
Sensitivity	.859 (.049) {.762 to .955}	.902 (.041) {.820 to .983}	.980 (.020) {.940 to .999}
Specificity	.770 (.060) {.651 to .889}	.816 (.056) {.704 to .928}	.808 (.057) {.696 to .920}
FN Error	.050 (.031) {.001 to .113}	.009 (.014) {.001 to .037}	.009 (.013) {.001 to .036}
FP Error	.126 (.046) {.035 to .217}	.091 (.041) {.009 to .172}	.181 (.056) {.071 to .292}
D Inc	.090 (.040) {.010 to .170}	.087 (.039) {.009 to .166}	.010 (.014) {.001 to .038}
T Inc	.102 (.042) {.018 to .186}	.092 (.042) {.009 to .176}	.009 (.013) {.001 to .035}
PPV	.870 (.047) {.778 to .962}	.909 (.04) {.829 to .989}	.843 (.047) {.750 to .935}
NPV	.939 (.038) {.864 to .999}	.987 (.017) {.954 to .999}	.989 (.016) {.956 to .999}
D Correct	.944 (.034) {.876 to .999}	.989 (.015) {.959 to .999}	.990 (.013) {.963 to .999}
T Correct	.858 (.051) {.757 to .959}	.899 (.045) {.810 to .988}	.816 (.056) {.705 to .928}

A series of two-way ANOVAs, model x case status, for correct decisions including inconclusives, errors, and inconclusive results showed that there were no statistically significant interactions and no significant main effect differences for the ESS and the automated ESS models, suggesting the automated ESS model replicates the manual ESS model[4].

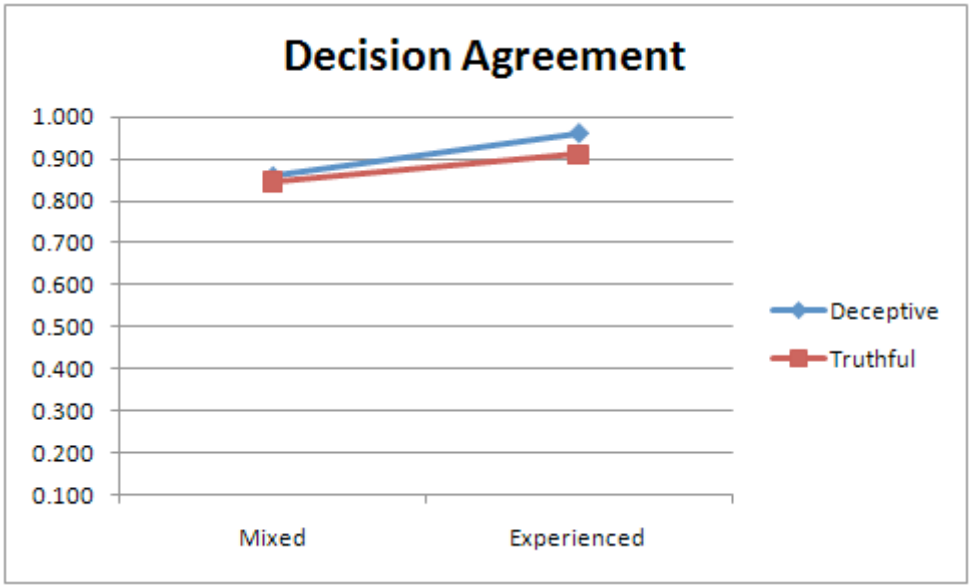
Interrater Reliability. A bootstrap of 1,000 iterations of the pairwise proportion of decision agreement, excluding inconclusive results, showed that the experienced cohort had higher overall rates of agreement than

4 The automated ESS model replicates the same procedures as the manual ESS model, with automated execution of Kircher measurements and automated execution of transformation and decision rules. Ratio comparison scores were made at 1.1:1 using the stronger adjacent comparison question. One noteworthy difference between the manual and automated models is the measurement of the pneumograph data. Automated measurements were made with the RLL feature, while manual scores were made using a simplified pattern recognition approach as described in previous works on the ESS. The proportion of pairwise decision agreement between manual and automated ESS models, reported in this study, indicates that this procedural difference is empirically meaningless.

the mixed experience cohort. The mixed-experience cohort had a mean rate of decision agreement of .853 (SD = .040), with 95% confidence range of .773 to .932. Decision agreement for deceptive cases was .860 (SD = .058) and agreement for truthful cases was .845 (SD = .063) for the mixed experience cohort. Pairwise agreement for the experienced cohort was higher, with a mean proportion of .938 (SD = .044), and 95% confidence range of .852 to .999. Decision agreement for deceptive cases was .961 (SD = .051) and agreement for truthful cases was .909 (SD = .081) for the experience cohort.

Two-way ANOVA, cohort x status, for the proportion of decision agreement revealed a statistically significant interaction for experience and case status ($F_{1,348} = 23.820$, $p < .001$), shown in Figure 4. A series of one-way post hoc ANOVAs showed there were no significant differences in the proportion of decision agreement for either case status or for experience.

Figure 4. Decision agreement.



Manual ESS scores for each of the 22 cases in the confirmed You-Phase sample were averaged for the 15 scorers. Results were calculated for the averaged scores, and those results were compared to the results of the automated ESS model and the OSS-3 computer algorithm. The proportion of pairwise decision agreement, excluding inconclusives, for the ESS, automated ESS, and OSS-3 algorithm was .988 (SD = .02), with a 95% confidence range of .948 to .999.

ESS and Seven-position Models. A series of two-way ANOVAs, model x status, was used to compare decision accuracy, errors and inconclusive results for the ESS and the seven position model using two-stage decision rules. There was no significant interaction for model and status and no significant main effect for decisions including inconclusives. Figure 5 shows the mean plots for decisions with inconclusives. There was a significant interaction of model and status for errors ($F_{1,546} = 212.49$, $p < .001$), shown in Figure 6. One way post hoc analysis showed there were no significant differences in error rates for the two seven-position and ESS models within the truthful and deceptive groups. There was a significant interaction of model and status, as shown in Figure 7, for inconclusive results ($F_{1,546} = 24.693$, $p < .001$). However, one-way post hoc AVOVAs found no significant differences in inconclusive rates within the truthful and deceptive groups. This suggests the features and transformations of the ESS and seven position federal models are capable of extracting similar diagnostic information. For these analyses the two-stage decision model was used to hold decisions constant and allow comparison of features and transformations.

Figure 5. Decisions including inconclusive results for seven-position and ESS models.

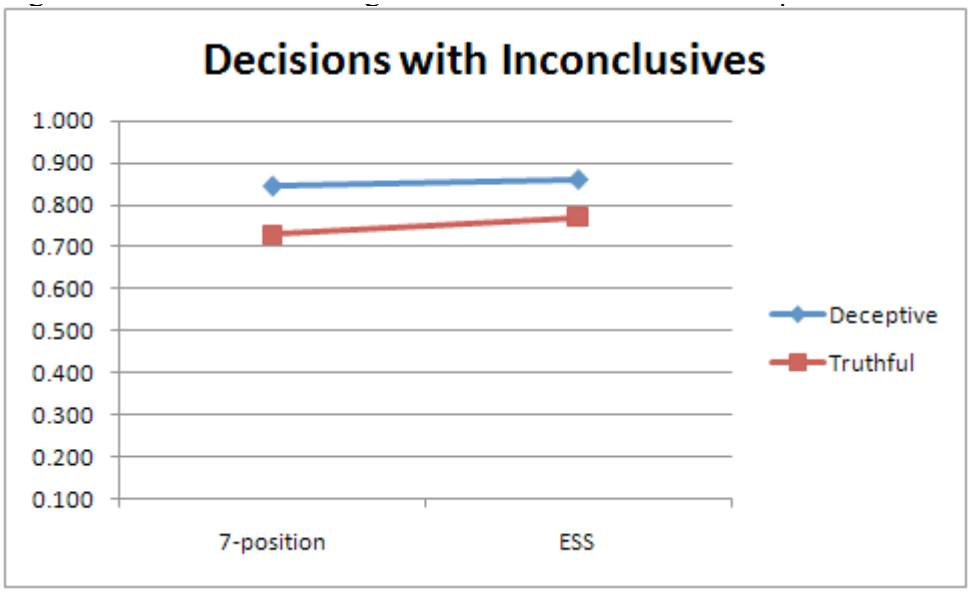


Figure 6. Decision errors for seven-position and ESS models.

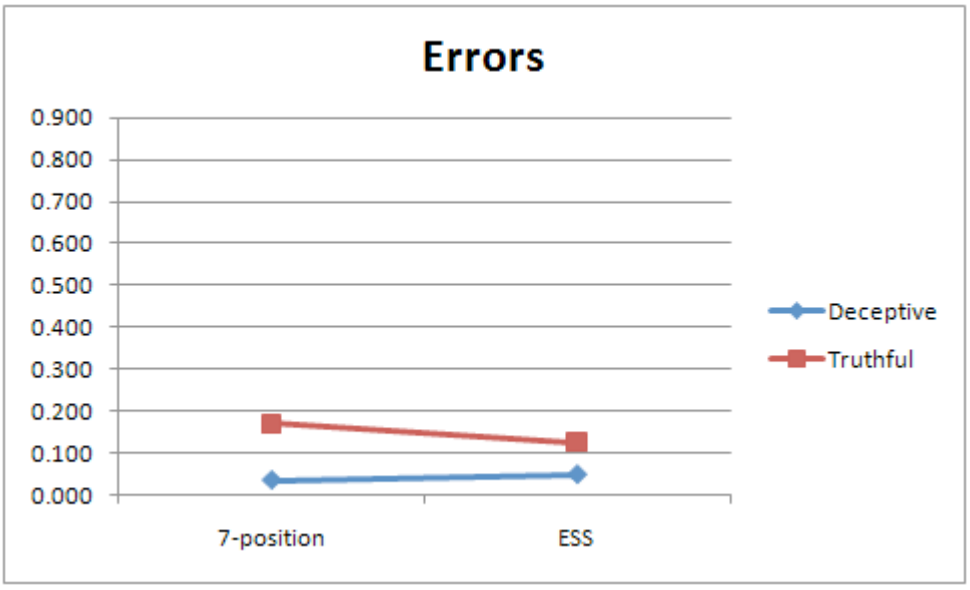
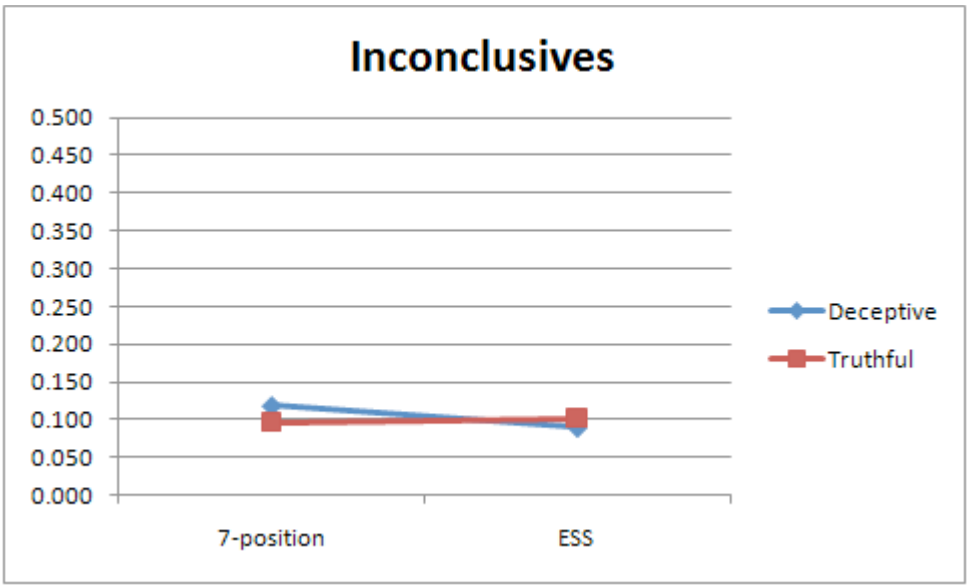


Figure 7. Inconclusive results for seven-position and ESS models.



Level of Experience. Additional analysis was completed to determine if there were differences resulting from the level of experience of the scorers. Table 3 shows the ESS accuracy profiles for 8 inexperienced scorers and 7 experienced scorers.

Table 3. Mean, standard deviation, and 95% confidence intervals for inexperienced and experienced scorers.

	Inexperienced Scorers	Experienced Scorers
Unweighted Accuracy	.859 (.036) {.788 to .930}	.955 (.021) {.913 to .997}
Unweighted Inc	.084 (.026) {.032 to .137}	.110 (.030) {.050 to .170}
Sensitivity	.827 (.053) {.723 to .931}	.896 (.043) {.810 to .982}
Specificity	.744 (.063) {.620 to .867}	.803 (.055) {.695 to .912}
FN Error	.070 (.037) {.001 to .143}	.024 (.021) {.001 to .067}
FP Error	.188 (.056) {.077 to .299}	.054 (.031) {.001 to .116}
D Inc	.101 (.041) {.020 to .183}	.079 (.037) {.005 to .152}
T Inc	.067 (.034) {.001 to .136}	.142 (.049) {.044 to .239}
PPV	.815 (.055) {.706 to .925}	.943 (.033) {.878 to .999}
NPV	.912 (.045) {.822 to .999}	.970 (.026) {.918 to .999}
D Correct	.921 (.041) {.839 to .999}	.973 (.024) {.926 to .999}
T Correct	.797 (.060) {.679 to .915}	.937 (.037) {.864 to .999}

A series of two-way ANOVAs, experience x case status, was calculated for the proportions of correct decisions including inconclusives, errors, and inconclusive results. The interaction of experience and case status was not significant for correct decisions including inconclusives, nor were the main effects, as shown in Figure 8. Figure 9 shows the significant interaction of experience and case status ($F_{1,326} = 83.928$, $p < .001$) for decision errors. One-way post hoc ANOVAs showed that the differences in decision errors were significant only for the false-positive errors ($F_{1,162} = 4.383$, $p = .038$). The inexperienced scorers produced more false-positive errors. The difference in false-negative errors was not significant. Figure 10 shows the significant interaction of experience and case status for inconclusive results ($F_{1,326} = 123.005$, $p < .001$). Experienced scorers produced more inconclusive results for truthful cases and fewer inconclusive results for deceptive cases compared to the inexperienced scorers. However, one-way post hoc ANOVAs revealed that the differences in inconclusive rates within the truthful and deceptive groups were not significant, nor was the difference in inconclusive rates for the combined deceptive and truthful cases.

Figure 8. Correct decisions for experienced and inexperienced scorers including inconclusives.

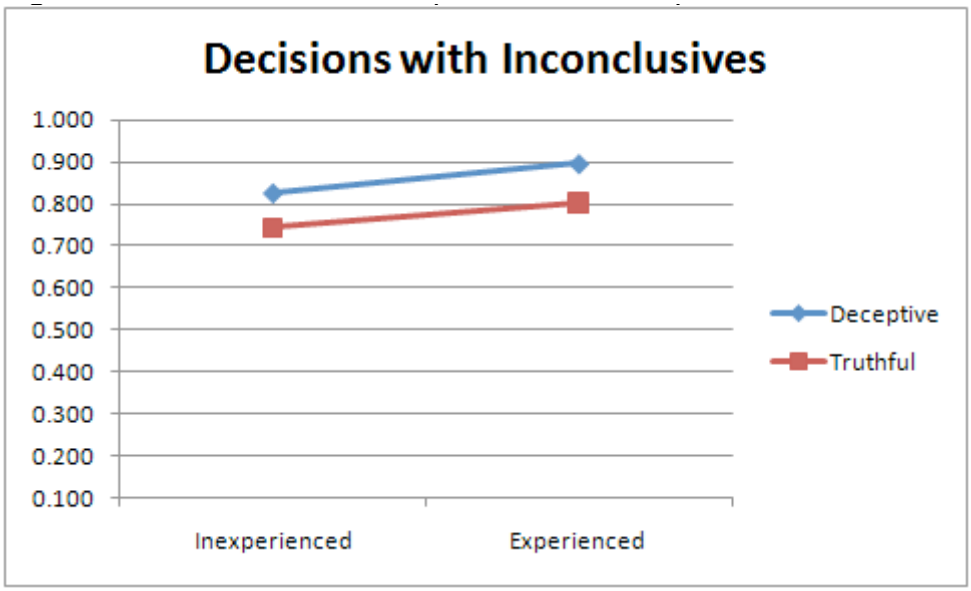


Figure 9. Decision errors for inexperienced and experienced scorers.

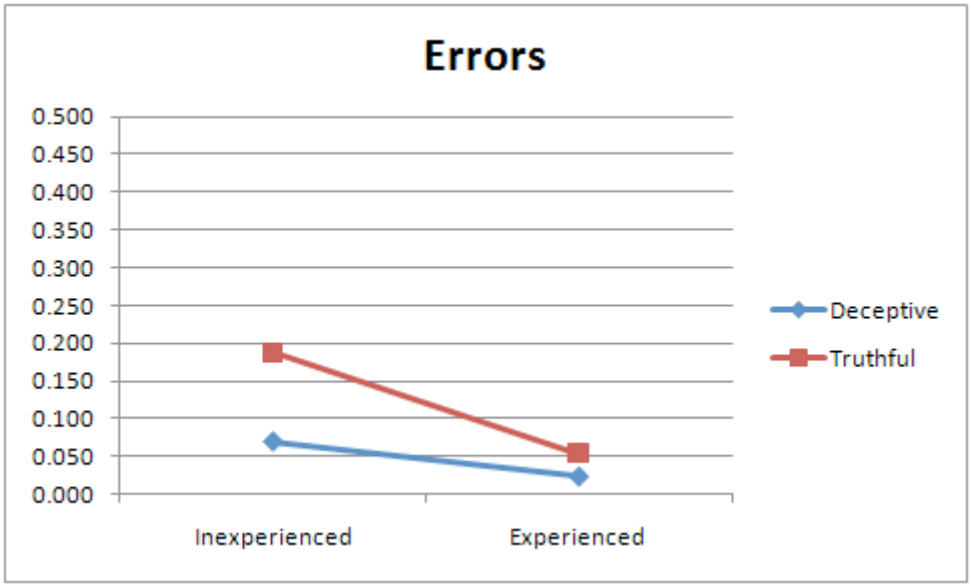
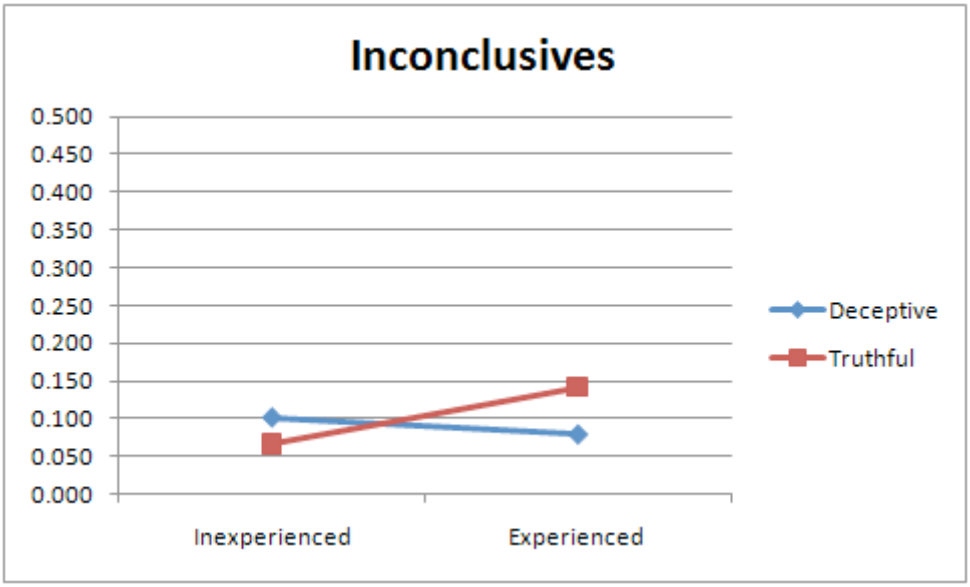


Figure 10. Inconclusive results for inexperienced and experienced scorers.



Discussion

The confirmed You-Phase examinations in this study produced overall criterion accuracy rates well above chance with all of the TDA models used. However, differences were observed among the dimensional profiles of criterion accuracy produced by the different TDA models. The three-position TDA model was unable to correctly identify truthful cases at rates greater than chance when scored with traditional decision rules and with two-stage decision rules. Additionally, inconclusive rates were highest for the three-position TDA model, exceeding 20% for both decision rules.

Criterion accuracy for the combined cohorts of experienced and inexperienced scorers exceeded 90% for ESS scores, with inconclusive results less than 10%. Interrater decision agreement, excluding inconclusive results, was high for both inexperienced scorers, exceeding 85% and experienced scorers, exceeding 95%.

Both the OSS-3 computer algorithm and the automated ESS model also produced results for the confirmed You-Phase examinations that were significantly more accurate than chance, for both deceptive and truthful cases. Results of this study further confirm the correspondence between ESS scores and ESS scores that are calculated from the seven-position federal transformation model. There were no significant interactions or main effect differences for the ESS and automated ESS models. These results support the potential for further automation of the ESS. Of course, no manual or automated TDA method can be expected to accurately interpret the results of a test that has been conducted improperly or ineffectively.

Experienced scorers outperformed inexperienced scorers using the ESS. However, differences in overall decision accuracy were not significant, nor were differences in overall error. Experienced scorers in this study produced significantly fewer false-positive errors than the inexperienced scorers. In other words, diagnostic efficiency and the ability to detect deception was not significantly different for the experienced and inexperienced scorers, but the experienced scorers misclassified fewer truthful cases. Future studies should continue to explore the role and value of experience in PDD administration and TDA.

Limitations of the present study include the small sample size, and the unknown degree to which the sample is representative of the population. We also note that one experienced scorer participated in both the mixed experience and experienced cohort using different scoring methods, introducing a source of shared variance for the two cohorts. We have attempted in our analysis to verify that the sampling distribution does not differ significantly from a Monte Carlo estimate of You-Phase sampling distributions, constructed from other available data. Another limitation to the present study is that we were unable to compare the experienced and inexperienced scorers using the seven-position and three-position TDA models. Another third obvious limitation is the lack of information about how the confirmed cases were selected for inclusion in the archive, other than the availability of the confirmation data, and the unknown possibility that the exclusion of field cases without confirmation data might lead to an overestimation of criterion accuracy rates.

Just as no single sample can adequately represent the population as a whole, no single study can be regarded as a definitive description of the potential accuracy of a PDD examination technique. Coupled with the results of others studies based on other sampling data, these results indicate that the You-Phase technique can be capable of providing criterion validity at a satisfactory level for evidentiary testing. Certainly no PDD format or TDA model can be expected to produce satisfactory results if the data collection is not based on effective administration and use of the testing procedures and testing instruments. As always, additional research is warranted to better understand and further confirm the capabilities of this test format.

Results of this study support the validity of the hypotheses that the You-Phase technique, when scored via the Federal seven position, ESS, or OSS-3 models, can differentiate confirmed deceptive from confirmed truthful field investigation cases at rates that that are significantly greater than chance. Although the rate of inclusive results was high for the three position scoring method, overall decision accuracy was not inconsistent with the other results. These results suggest continued interest in the You-Phase technique as an effective diagnostic test format in field settings. These results further suggest continued interest in the ESS and 7-position models for manual TDA, and the OSS-3 and automated ESS models for automated TDA, with the obvious caveat that it is not realistic to expect perfect or near-perfect accuracy in field settings. Any suggestion that any PDD technique or TDA model can provide near-perfect accuracy in field settings should be viewed with great caution, and should be subject to intense scrutiny of the supporting data before being accepted.

Acknowledgements

We are extremely grateful to Pat O'Burke, Chip Morgan, 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, Nancy Sley, and Eric Alfonso. Without the commitment of these professionals none of this work would have been accomplished.

References

Backster, C. (1963). Standardized polygraph notepack and technique guide: Backster zone comparison technique. Cleve Backster: New York.

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 (2006a). Test Data Analysis:DoDPI numerical evaluation scoring system. Retreived from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007.

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

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

Honts, C. R. & Hodes, R.L. (1982). The effects of multiple physical countermeasures on the detection of deception. Psychophysiology, 19, 564-565. (abstract)

Honts, C. R. & Hodes, R.L. (1983). The detection of physical countermeasures. Polygraph, 12, 7-17.

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.

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

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

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

Meiron, E., Krapohl, D. J. & Ashkenazi, T. (2008). An assessment of the Backster “Either-Or” Rule in polygraph scoring. Polygraph, 37, 240-249.

Nelson, R. (2012). Monte Carlo study of criterion validity for two-question Zone Comparison Tests with the Empirical Scoring System, seven-position and three-position scoring models.

Nelson, R. & Blalock, B. (2016). 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 45(1), in press.

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

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, in press.

Nelson, R., Blalock, B., Oelrich, M. & Cushman, B. (2011a). Reliability of the Empirical Scoring System with expert examiners. Polygraph, 40, (in press).

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

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.

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

Senter, S. M. (2003). Modified general question test decision rule exploration. Polygraph, 32, 251-263.

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

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