Monte Carlo Study of Criterion Validity for Two-Question Zone Comparison Tests with the Empirical Scoring System, Seven-Position and Three-Position Scoring Models

Raymond Nelson

Abstract

Monte Carlo methods were used to estimate criterion accuracy levels for PDD examinations conducted with the two-question ZCT format, also known as the You-Phase technique and the Bi-Zone technique. Results show that PDD examinations conducted with the You-Phase technique, can be expected to meet or exceed 90% mean decision accuracy with fewer than 20% mean inconclusives using the ESS and seven-position federal TDA models. There were no significant differences in unweighted decision accuracy for the ESS, seven-position and three-position TDA models. Three-position TDA resulted in increases in inconclusive results over 20%. ANOVAs showed no significant differences in test sensitivity to deception for the three TDA models. There were significant differences in: decreased test specificity and increased inconclusives for truthful cases with the seven-position and three-position models using traditional decision rules. These results suggest that seven-position and ESS numerical transformations are capable of extracting similarly useful diagnostic information from the PDD test data. Continued interest in the two-question ZCT format is recommended along with the ESS and seven-position TDA models. Comparison of the Monte Carlo results with field and laboratory data is recommended.

Introduction

The You-Phase technique, sometimes referred to as the Bi-Zone Technique, is a single-issue diagnostic format for Psychophysiological Detection of Detection (PDD) tests. The You-Phase technique is part of the family of Zone Comparison Techniques (ZCT), all of which emanate from the work of Backster (1963) and Reid (1947). You-Phase examinations consist of three comparison questions and two investigation target (relevant) questions that describe the examinee's behavioral involvement in a single known incident or allegation. Other procedural questions are also used. The name of the You-Phase technique is a reference to the basic robust form of the relevant stimulus question, "Did you do it?"

Two closely related versions of the You-Phase technique exist today: a version taught by the Department of Defense (2006) and several polygraph schools accredited by the American Polygraph Association, and the version originally developed by Backster. These two versions differ primarily in their method of test data analysis, including features, transformation rules, decision rules, and cutscores. There are no substantive differences in the sequence of test questions or principles for target selection and question formulation for these two versions. The Backster version of the You-Phase technique

Raymond Nelson is a research specialist with the Lafayette Instrument Company (LIC) and an elected member of the APA Board. The views expressed in this work are those of the author and do not necessarily represent those of the LIC or the APA. Mr. Nelson is a psychotherapist, polygraph field examiner, developer of the OSS-3 scoring algorithm, and is the author of several publications on various polygraph topics. 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.

Special thanks to Mark Handler and Benjamin Blalock who read early drafts of this manuscript, and to Pam Shaw and Donald Krapohl who reviewed and critiqued it prior to publication.

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 and Ashkenazi (2008) studied the Backster "either-or" rule used with the Backster You-Phase technique. These studies, however, were not designed to address the issue of criterion accuracy. The present study is limited to the Federal version of the You-Phase technique.

Although the You-Phase technique is supported by a complete procedural description (Department of Defense, 2006), and by favorable opinions anchored in decades of case experience and anecdotal evidence, there are no published studies that describe the criterion accuracy of this technique with the evidence-based TDA models that now exist, including the seven-position and three-position systems (Department of Defense, 2006), the ESS (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Krapohl, 2010; Nelson & Krapohl, 2011; Nelson, Blalock, Oelrich & Cushman, in press; Nelson, Krapohl & Handler, 2008), or the Objective Scoring System, version 3 (OSS-3) (Nelson et al., 2008). This study was designed to investigate the hypothesis that two-question ZCT examinations can discriminate deception from truth-telling at rates greater than chance.

Method

Bootstrap Monte Carlo[1] methods were used to develop normative parameters that would be used to calculate the level of statistical significance for each test result in a second Monte Carlo model designed to study the criterion validity of the You-Phase technique.

Normative parameters for truthful and deceptive groups were calculated from a Monte Carlo space consisting of 100 simulated two-question ZCT examinations for which the deceptive and truthful scores were resampled, for each iteration of the Monte Carlo space, from the subtotal scores from first two relevant questions from the seven participants in the Nelson et al. (2008) study. In accordance with the target selection and question formulation principles set forth for the You-Phase technique (Department of Defense, 2006), these question scores pertain to direct involvement in a single issue of concern. A single non-independent criterion status was set for each case in the Monte Carlo space by comparing a random number to a fixed base rate of .5. Another random number was standardized to the resampled deceptive or truthful normative parameters according to the status of each case in the Monte Carlo space. The Monte Carlo space was recalculated for 10,000 iterations. Resampling of the deceptive and truthful seed scores, and resulting parameters, for each iteration of the Monte Carlo space was intended to increase uncontrolled Monte Carlo variance of the resulting normative parameters. Resampled distributions of mean and standard deviation statistics were normally distributed, indicating that the distributions of two-question ZCT scores can also be expected to be normally distributed. Resulting Monte Carlo normative mean and standard deviation statistics are therefore descriptive of a normal distribution whose parameters are the bootstrap means of the mean and standard deviation of the ESS scores for the first two relevant questions scores provided by the seven participants in the Nelson et al. (2008) study.

A second Monte Carlo model was then used to calculate the criterion accuracy profile for two-question ZCT examinations. The second Monte Carlo space consisted of 60 two-question ZCT exams for which relevant question subtotal scores were simulated using seed data in the form of subtotal scores from an archival sample of confirmed field examination cases (N = 60) that was scored by six experienced examiners in a previous study on the ESS and seven-positions TDA with three-question ZCT examinations (Nelson &

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

Krapohl, 2011). Thirty of the seeding cases were confirmed as deceptive, and the other 30 were confirmed as truthful. The first two questions in the three-question ZCT examinations pertain to direct involvement in the issue under investigation, and correspond to the questions in the two-question ZCT format. Deceptive and truthful subtotal scores from the seeding cases were then randomly selected according to the case status. The criterion status of each simulated case was set by comparing a random number to the base-rate of .5. Although a single criterion was set for each simulated exam in the Monte Carlo space, the subtotal scores were selected independently from within deceptive and truthful cases, meaning that seeded subtotal scores for each simulated case were randomly selected from different cases in the seed data.

Three versions of the second Monte Carlo model were created: one to simulate You-Phase results with the ESS, a second to simulate results of the seven-position TDA model, and a third to simulate results from the three-position TDA model. Two additional sub-versions of the Monte Carlo models were created to simulate seven-position and three-position results when scored using two-stage decision rules (Krapohl, 2005; Krapohl & Cushman, 2006; Senter, 2003; Senter & Dollins, 2004, 2008a, 2008b). Two-stage rules can be thought of as a combination of the grand total rule and the Spot Score Rule (SSR) (Light, 1999) in which the robust effectiveness of the grand total is used at stage one, while the subtotals are used to reduce inconclusives and increase test sensitivity at stage two if the result from the grand total is inconclusive. Use of the SSR in this manner serves to reduce the occurrence of inconclusive results and increase test sensitivity to deception without creating a concurrent large increase in false-positive errors. Two-stage rules achieve a procedural approximation of a Bonferonni correction to protect against inflated alpha and increased false-positive errors when conducting multiple significance comparisons for a single classification. Each of the Monte Carlo models was run for 1,000 iterations.

Alpha was set at .1 for truthful classifications and .05 for deceptive classifications with the ESS. These alpha boundaries correspond to grand total cutscores of +2 or greater for truthful classifications and -4 or lower for deceptive classifications. To prevent any increase in false-positive errors as a result of inflation of the alpha when making multiple statistical comparisons for a single examination issue, decisions based on subtotals were made using a Bonferonni correction to the desired alpha (.05 / 2 = .025). The Bonferonni corrected alpha corresponded to a subtotal cutscore of -6 or lower for deceptive classifications based on subtotal scores.

Cutscores for the seven-position and three-position models were those described by the Department of Defense (2006): grand total of -4 or lower or any subtotal of -3 or lower for deceptive classifications, and a grand total of +4 or greater and all subtotals greater than 0 for truthful classifications. To further evaluate the effectiveness of the seven-position and three-position models, both systems were evaluated using two-stage decision rules (Krapohl, 2005; Krapohl & Cushman, 2006; Senter, 2003; Senter & Dollins, 2004, 2008a, 2008b). Cutscores for the two-stage decision rules were a grand total of +4 or greater for truthful classifications or a grand total of -4 or lower for deceptive classifications at stage one, and any subtotal of -3 or lower for deceptive classifications at stage two. There are no truthful classifications at stage two when using two-stage decision rules. Cutscores for two-stage rules for the three-position TDA model were the same as those for the seven-position model.

Results

All statistical analyses were completed with a level of significance set at alpha = .05.

Reliability of the Monte Carlo Model

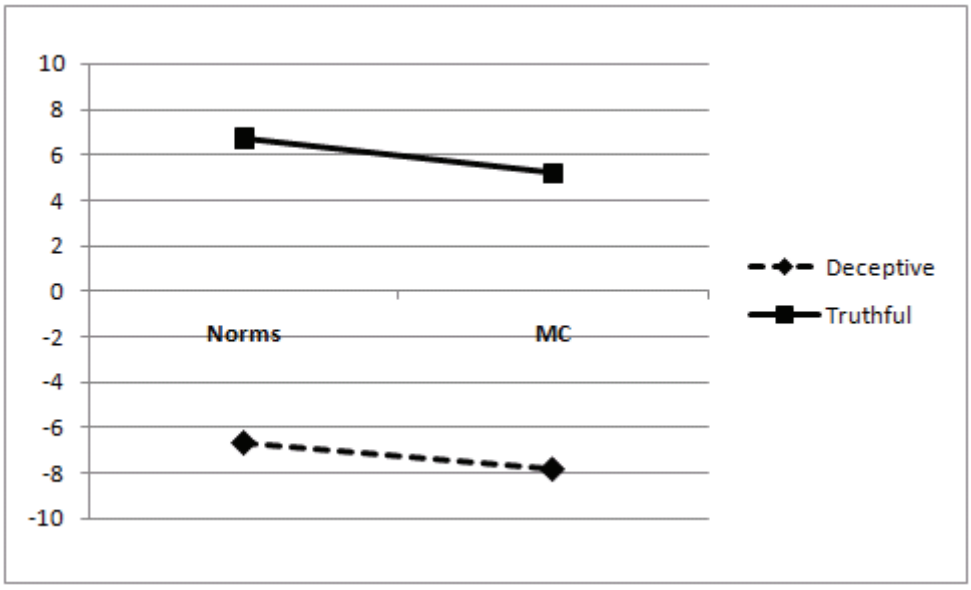
The normative Monte Carlo mean for deceptive ESS total scores of two-question ZCT normative simulation was -6.685 and the Monte Carlo standard deviation for deceptive cases was 6.881. The Monte Carlo mean for truthful cases was 6.735 and the Monte Carlo standard deviation was 6.045. Because field PDD examinations are scored in integers, not real numbers, normative parameters were truncated to integers. See Appendix A for a table of normative data for two-question ZCT examinations.

The second Monte Carlo You-Phase simulation produced a mean total deceptive score of -7.859 (SD = 4.937), and a mean total truthful score of 5.188 (SD = 5.926). A two-way ANOVA was used to compare the total scores of the second Monte Carlo model with the normative parameters developed with the first Monte Carlo. There were no significant main effect differences for the absolute value of total scores, and no significant interaction effects (See Table 1). However, the interaction between model and status was approaching a significant level ($p < 0.11$). Figure 1 shows the mean plot for the two Monte Carlo You-Phase models.

Table 1. Two-way ANOVA summary for Monte Carlo (MC) norm and absolute total scores.

Source	SS	df	MS	F	p	F crit .05
Model	1.735	1	0.017	0.000	0.982	3.889
Status	85.875	1	0.859	0.024	0.877	3.889
Interaction	92.553	1	92.553	2.582	0.110	3.889
Error	7025.525	196	35.845			
Total	180.163	199				

Figure 1. Mean plot for two You-Phase Monte Carlo (MC) models.



For the seven-position You-Phase Monte Carlo model, the mean total score for deceptive cases was -6.398 (SD = 4.914), and the mean total for truthful cases was 5.485 (SD = 5.106). Monte Carlo mean totals for the three-position You-Phase model were -4.720 (SD = 3.345) for deceptive cases and 3.901 (SD = 3.804) for deceptive cases.

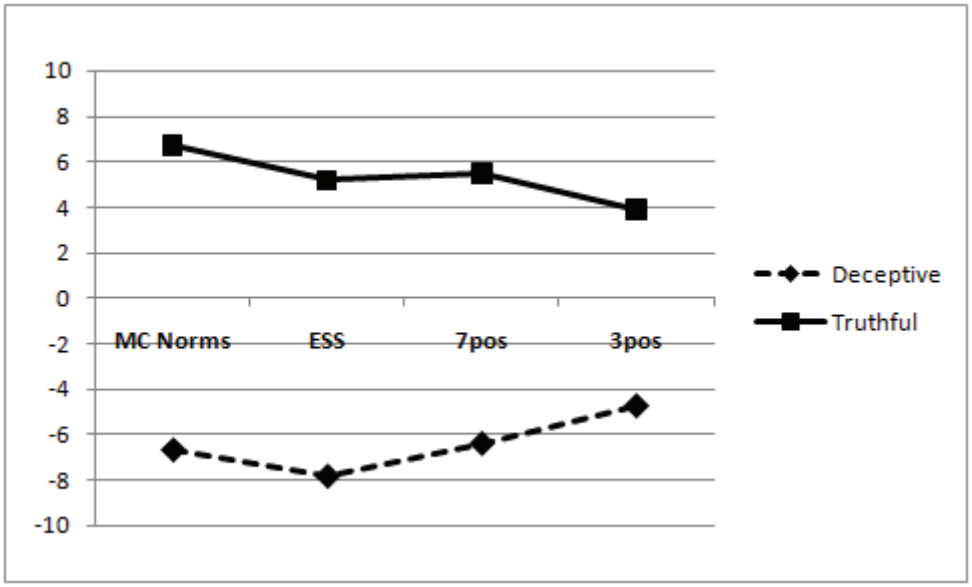
Two-by-four ANOVA comparisons of the Monte Carlo means for the seven-position and three-position models resulted in a

significant interaction between transformation model and case status, shown in Table 2 and Figure 2. A series of ANOVA contrasts revealed that three-position models differed from each of the others at or below the .01 level. That difference was related to weaker absolute scores and the loss of diagnostic information for the three-position model.

Table 2. Two-way ANOVA summary for transformation models.

Source	SS	df	MS	F	p	F crit .05
Model	357.050	3	3.571	0.131	0.942	2.628
Status	118.399	1	0.592	0.022	0.883	3.865
Interaction	97.612	1	97.612	3.576	0.049	3.865
Error	10699.169	392	27.294			
Total	573.061	397				

Figure 2. Mean total scores of Monte Carlo (MC) norms, ESS, seven-position and three-position transformation models.



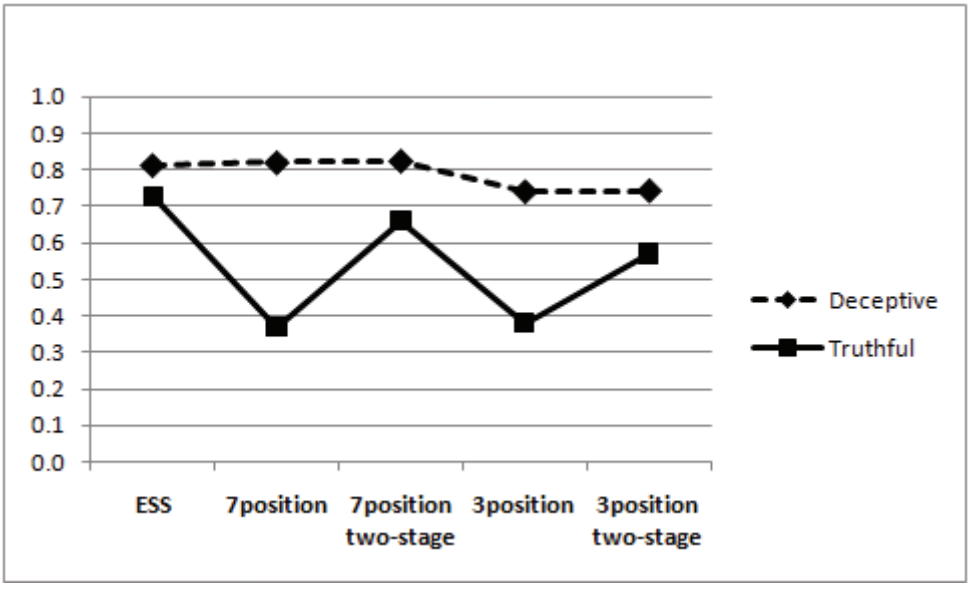
Criterion Accuracy

Criterion accuracy profiles were calculated, including mean, standard deviations and statistical confidence intervals for sensitivity to deception, specificity to truthfulness, inconclusive results for deceptive and truthful cases, false-positive and false-negative errors, positive predictive value (PPV), negative predictive value (NPV), the proportion of correct decisions without inconclusives for truthful and deceptive cases, and the unweighted mean of correct decisions and inconclusives results. Table 3 shows the Monte Carlo accuracy profiles for two-question ZCT examinations for the ESS, seven-position and three-position TDA models.

Table 3. Mean, (standard deviation), and {95% confidence intervals} for criterion accuracy.

(N = 66)	ESS	7-position	7-position two-stage rules	3-position	3-position two-stage rules
Unweighted Accuracy	.920 (.029) {.863 to .978}	.870 (.042) {.788 to .952}	.913 (.031) {.852 to .974}	.890 (.044) {.804 to .975}	.921 (.031) {.861 to .984}
Unweighted Inconclusives	.159 (.041) {.078 to .239}	.229 (.044) {.143 to .315}	.169 (.041) {.088 to .25}	.324 (.051) {.224 to .424}	.273 (.050) {.174 to .372}
Deceptive Inconclusive	.154 (.051) {.054 to .254}	.157 (.054) {.052 to .262}	.154 (.052) {.053 to .256}	.260 (.064) {.135 to .385}	.255 (.064) {.129 to .381}
Truthful Inconclusive	.171 (.055) {.064 to .278}	.444 (.071) {.305 to .583}	.214 (.06) {.096 to .332}	.513 (.074) {.368 to .658}	.327 (.066) {.197 to .456}
Sensitivity	.813 (.055) {.705 to .920}	.833 (.055) {.725 to .941}	.829 (.054) {.724 to .935}	.740 (.064) {.615 to .865}	.743 (.065) {.616 to .869}
Specificity	.729 (.066) {.600 to .858}	.417 (.068) {.284 to .551}	.664 (.068) {.531 to .798}	.380 (.071) {.240 to .519}	.570 (.069) {.436 to .705}
False-negative	.033 (.025) {.001 to .083}	.010 (.014) {.001 to .038}	.016 (.018) {.001 to .052}	.001 (.001) {.001 to .001}	.002 (.007) {.001 to .015}
False-positive	.099 (.042) {.017 to .182}	.138 (.049) {.042 to .234}	.111 (.047) {.030 to .213}	.107 (.045) {.019 to .196}	.103 (.043) {.018 to .188}
PPV	.892 (.045) {.804 to .980}	.857 (.051) {.757 to .956}	.873 (.049) {.777 to .968}	.872 (.052) {.770 to .975}	.878 (.050) {.780 to .976}
NPV	.956 (.033) {.891 to .999}	.977 (.033) {.913 to .999}	.976 (.027) {.924 to .999}	.999 (.001) {.999 to .999}	.996 (.012) {.973 to .999}
Deceptive Correct	.961 (.030) {.902 to .999}	.988 (.017) {.954 to .999}	.981 (.021) {.939 to .999}	999 (.001) {.999 to .999}	.997 (.009) {.980 to .999}
Truthful Correct	.880 (.051) {.780 to .980}	.752 (.081) {.593 to .910}	.846 (.058) {.731 to .96}	.779 (.087) {.609 to .950}	.847 (.062) {.726 to .969}

Figure 3. Sensitivity and specificity for You-Phase TDA models.



Multivariate analyses revealed significant two-way (model x case status) interaction effects for correct decisions with inconclusives, also known as sensitivity and specificity [F (1,290) = 344.518 (p < .001)], decision errors [F (1,290) = 9.720 (p < .002)], and inconclusive results [F (1,290) = 280.821 (p < .001)]. A series of post hoc one-way ANOVAs showed there were no significant differences in the TDA models for sensitivity to deception or false-negative errors. However, differences in inconclusives for deceptive cases was significant [F (4,183) = 2.510 (p < .05)]. Differences between the TDA models were also significant for test specificity to truth-telling [F (4,183) = 14.910 (p < .001)], and inconclusive results for truthful cases [F (4,183) = 14.965 (p < .001)]. These differences appear to be related to the use of traditional decision rules, and are illustrated by the sensitivity and specificity levels shown in Figure 3. There were no significant one-way effects for false-negative or false-positive errors.

Discussion

This study used data from two different archival samples of ZCT exams to develop Monte Carlo models to study the criterion accuracy of the You-Phase technique. The first archival sample was used to develop Monte Carlo norms that were used to calculate the level of statistical significance that was used to make classification decisions for individual exams in the Monte Carlo model. A second archival sample was used to model two-question ZCT examination results, which were evaluated and classified using Monte Carlo norms from the first archival sample. Three different Monte Carlo simulations were constructed using the second archival sample, for the ESS, seven-position, and three-position TDA models, and the results were analyzed for each.

Results from this study indicate the You-Phase technique can discriminate confirmed truthful from confirmed deceptive cases at rates that are significantly greater than chance (p < .001), with no significant differences between the different TDA models in terms of unweighted decision accuracy, PPV or NPV, or the proportion of correct decisions for deceptive and truthful cases. Observed differences in Table 1 are attributable to uncontrolled variance within the archival sample data, not accounted for by the TDA models, or to Monte Carlo variance.

Although there were no significant differences in the performance of the TDA models with deceptive cases, there were significant differences in the abilities of the TDA models to make decisions and avoid inconclusives with the You-Phase exams. The three-position scoring model produced excessive inconclusives, as did the traditional seven-position model. For the three-position system, the deficiency appears to be related to both the loss of diagnostic information with three-position transformations, and to the use of suboptimal decision rules. For the seven-position model, excessive inconclusives appear to be attributable only to the decision rules and not to the numerical transformations. This should be the focus of continued research.

There were significant differences in the performance of the TDA models with truthful cases in the You-Phase simulations. Seven-position and three-position models showed significantly weaker test specificity, and greater inconclusives for truthful cases. These differences appear to be primarily influenced by the decision rules and cutscores and not the scoring features or numerical transformations, a finding consistent with those of Senter (2003), and Senter and Dollins (2004; 2008a; 2008b). These results suggest that ESS and seven-position numerical transformations may extract the same diagnostic content from the test data, and that three-position transformations are more blunted. This should be explored more fully in future research.

Although overall decision accuracy does not differ significantly for You-Phase examinations with the different TDA models, the observed differences may be important as both a practical and an ethical matter. It may be difficult to justify the use of suboptimal TDA models, in settings where human decisions may be influenced by the test outcome, when an expedient solution exists that will allow for control over error rates while constraining inconclusive results to manageable levels.

The ESS offers advantages that the other models do not. These advantages include a foundation of empirical support for all procedures and assumptions, and the ability to calculate and control both false-positive and false-negative errors according to operational priorities. Another practical advantage of a simple evidence-based TDA model is that of increased ease of skill acquisition and skill retention, both of which relate to increased interrater reliability. This should be investigated further in future studies.

Limitations of this study primarily involve the study design as a theoretical project using existing data to simulate and study the decision theoretic model of You-Phase examinations. Monte Carlo models are not intended to solve all problems or answer all questions, but are a highly useful way of using existing knowledge and extant data for studying complex and abstract problems for which it is impracticable or very difficult to accomplish in other ways. It should go without saying that Monte Carlo and theoretical studies should always be accompanied by field and laboratory experiments.

In conclusion, these results support the validity of the hypothesis that the You-Phase technique can discriminate confirmed deceptive from confirmed truthful cases at rates that are significantly greater than chance. Continued interest in the You-Phase technique is recommended along with continued interest in ESS and seven-position TDA models.

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 (2006). *Federal psychophysiological detection of deception examiner handbook*. Reprinted in *Polygraph*, 40(1), 2-66.

Handler, M., Nelson, R., Goodson, W. & Hicks, M. (2010). Empirical scoring system: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph*, 39, 200-215.

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

Honts, C. R. & Hodes, R.L. (1983). The detection of physical countermeasures. *Polygraph*, 12(1), 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. J. (2005). Polygraph decision rules for evidentiary and paired testing (Marin Protocol) applications. *Polygraph*, 34, 184-192.

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

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

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.

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. (2004). Comparison of question series and decision rules: A replication. *Polygraph*, 33, 223-233.

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

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

Appendix A

Monte Carlo norms for You-Phase examinations with the Empirical Scoring System

Deceptive Mean = -6.685 (SD = 6.881)
Truthful Mean = 6.735 (SD = 6.045)

Parameters were truncated to integer scores +6 and -6 to produce the following lookup table.

You-Phase			
Truthful Lookup Table (based on the normative distribution of deceptive scores)		Deceptive Lookup Table (based on the normative distribution of deceptive scores)	
Cutscore	p-value (alpha)	Cutscore	p-value (alpha)
-7	.570	7	.566
-6	.500	6	.500
-5	.434	5	.434
-4	.369	4	.369
-3	.309	3	.309
-2	.252	2	.252
-1	.202	1	.202
0	.159	0	.159
1	.122	-1	.122
2	.091	-2	.091
3	.067	-3	.067
4	.048	-4	.048
5	.033	-5	.033
6	.023	-6	.023
7	.015	-7	.015
8	.010	-8	.010
9	.006	-9	.006
10	.004	-10	.004
11	.002	-11	.002
12	.001	-12	.001
13	<.001	-13	<.001