Monte Carlo Study of Criterion Validity of Backster You-Phase Examinations

Raymond Nelson

Abstract

Monte Carlo methods were used to calculate a dimensional profile of criterion accuracy for PDD examinations conducted with the Backster You-Phase technique. Results show that decisions based on Backster You-Phase exams can be expected to discriminate deception from truth-telling at rates that are significantly greater than chance. Recommended cutscores for decisions based on two-charts outperformed results from decisions based on recommended cutscores for decisions based on three charts. Recommendations are made for future research regarding normative data, cutscores, and decision rules.

Introduction

The Backster You-Phase technique is an event-specific single-issue comparison question format for psychophysiological detection of deception (PDD) examinations, and is the well-spring from which several other single-issue PDD examination techniques have emerged. Both generic (Department of Defense, 2006) and boutique modifications have emerged (Gordon et al., 2005; Matte, 1978; Matte & Reuss, 1989) from the method first described by Backster. Although Reid (1947) provided the first description of a comparison question format, Backster (1963) provided the first highly standardized rationale and structure for the administration and scoring of a comparison question technique (CQT).

Two versions of the You-Phase technique exist today: the US Federal You-Phase format, taught by the US Department of Defense (2006), and the version originally developed by Backster (1963). These two versions differ in their scoring features, transformation rules, decision rules, and cutscores. The Federal You-Phase technique is scored with features similar to those developed at the University of Utah (Bell, Raskin, Honts & Kircher, 1999; Kircher, Kristjiansson, Gardner & Webb, 2005; Podlesny & Raskin, 1978; Raskin & Hare, 1978; Raskin, Kircher, Honts & Horowitz, 1988), while the Backster You-Phase technique is scored using physiological features defined by Backster, as described by Matte (1996) and Weaver (1980).

Esoteric differences may exist in linguistics and semantics or the Sacrifice Question, Symptomatic Questions, Relevant Questions (RQs) and Comparison Questions (CQs). These subtle differences may be argued by proponents and adherents of various PDD techniques as related to differences in criterion accuracy. Capps (1991) and Horvath (1994) showed evidence that the Sacrifice Relevant question does not serve to absorb initial responsivity and does not protect against errors, as hypothesized. Although Backster (2001) and Matte (2001) have voiced their opinions in support of the symptomatic hypothesis, evidence so far has suggested that the symptomatic question also does not function as intended (Honts, Amato & Gordon, 2000, 2004; Krapohl & Ryan, 2001). Other research has failed to support the validity of hypotheses regarding the contribution of linguistic and structural differences in CQs to criterion accuracy (Amsel, 1999; Horvath, 1988, 1991; Horvath & Palmatier, 2008; Palmatier, 1991). In general, research has not supported the construct validity or structural validity of technical questions based on esoteric linguistics, and attempts to exploit the precision of verbal logic, as a contributor to PDD test accuracy. However, the abundance of studies describing comparison question PDD test accuracy as significantly greater than chance does support the construct validity of the general categories of CQs and RQs.

The sequence of test questions for the Federal and Backster You-Phase formats can be seen in Table 1. The order of questions differs slightly in that the order of the second and third test questions is reversed for the two You-Phase formats.

Table 1. Backster and Federal You-Phase formats.

	Backster You-Phase	Federal-You-Phase
1.	Neutral Question	Neutral Question
2.	Symptomatic Question	Sacrifice Relevant Question
3.	Sacrifice Relevant Question	Symptomatic Question
4.	Comparison Question	Comparison Question
5.	Relevant Question	Relevant Question
6.	Comparison Question	Comparison Question
7.	Relevant Question	Relevant Question
8.	Comparison Question	Comparison Question
9.	Symptomatic Question	Symptomatic Question

Responses to RQs of Federal You-Phase examinations are compared to responses of the strongest of reactions to the nearest CQs, using the seven-position and three-position models, and test data analysis rules described by the Department of Defense (2006).[1] Reactions to RQs of Backster You-Phase examinations are compared to the weaker of reactions to nearby CQs, using the Either-Or-Rule, unless the magnitude of response to the stronger of nearby CQs produces a linear ratio of 4:1 or greater, using the Green-Zone-Abuse-Rule. Numerical scores of Backster You-Phase examinations are assigned using a seven-position rubric based on linear ratios[2] and a system of 21 rules. A complete description of the Backster test data analysis rules is beyond the scope of this paper, but a list of rules can be seen in Appendix A.

Decision rules for the Federal You-Phase technique involve the combination of the grand total and subtotal scores, while decision rules for Backster You-Phase examinations use only the grand total score. Statistical descriptions of normative data have not been published for either the Federal You-Phase technique or the Backster You-Phase technique, and cutscores have been a matter of administrative policy not based on statistical probability distributions. Cutscores for the Federal You-Phase technique are +/- 4 for the grand total and +/-3 for subtotal scores of three to five presentations of the test question sequence. Cutscores for the Backster You-Phase technique, using the grand total score only, are + 5 or -9 for two charts, and +7 or -13 for three charts.

The Backster You-Phase technique was used by Honts et al., (1985) in a series of countermeasure studies, also reported in Honts and Hodes, (1983), and by Meiron et al. (2008) who studied the Either-Or-Rule. Neither of these studies is satisfactory as a study of generalizable criterion validity of the Backster You-Phase technique.

[1] Studies have also shown the effectiveness of scoring Federal You-Phase exams with the Empirical Scoring System (Nelson, In press; Nelson, Handler, Blalock & Cushman, In press).

[2] Handler et al (2010) described that electrodermal responses, like many human physiological responses, are non-linear, and often log-linear, and that linear assumptions are false regarding electrodermal reactions recorded during PDD examinations. Nevertheless, linear ratios have been used traditionally in PDD test data analysis models.

Meiron et al. (2008) studied a highly-selective, non-random, and non-representative sample of Backster You-Phase Exams (N = 100) to study the Backster Either-Or-Rule. Examinations with erroneous or inconclusive results were excluded from that sample,[3] and the results from the original examiners for the Meiron et al. (2008) cases included no errors. Unless one endorses the naïve belief that the PDD examination can provide perfect or near perfect accuracy, the Meiron et al. (2008) sampling distribution must be considered non-representative of the population of field cases as it includes diagnostic variance but is systematically devoid of error variance. Because there is no such thing as a perfect test, scientific studies of criterion validity require samples that contain normal proportions of diagnostic and error variance, making it impossible to justify the independent use of the Meiron et al. (2008) sample as a study of criterion accuracy and error rates. Indeed, the stated purpose of the study was not that of a criterion study, but an exploration of the role and contribution of the Backster Either-Or-Rule. The common method of ensuring the representativeness and generalizability of study data, and the replicability of study results, is to construct study samples through randomization while refraining from paring or manipulating the samples. Field PDD research samples are inherently non-random and potentially problematic in that cases are typically selected through the non-random determinant of the availability of confirmation data. Because error cases will lack confirmation data, field samples are regarded as at-risk for systematically excluding error variance and therefore capable of overestimating criterion accuracy. Replication and comparison of sample distribution parameters with those of other sampling distributions is a necessary part of any assertion of the representativeness of field sample data.

Laboratory samples, though more readily compliant with assumptions and requirements regarding randomization, are subject to different limitations from field samples, and have assumed though unknown ecological validity. An obvious threat to ecological validity for the Honts et al. (1985) sampling distribution (N = 48) is that the testing equipment for that study did not include a standard blood pressure cuff, and instead employed an alternative technology for measuring changes in cardiovascular activity. Although the variance of cardiograph scores in the sampling distributions of scores from the Honts et al. (1985) study may vary in unknown ways from the distributions of scores conducted using standard field testing equipment, pneumograph and electrodermal data for the Honts et al. (1985) study were obtained using sensors identical to those used in field settings. Despite the use of standard field practice sensors to record electrodermal and pneumograph data, attempts to portray the Honts et al. (1985) study as representative of field practices are not justified. The Honts et al. (1985) study produced criterion results that were imperfect though significantly greater than chance.

The present study is intended to help fill a gap in the published literature regarding criterion accuracy of the Backster numerical scoring system and the You-Phase technique. The hypothesis was that numerical scores of You-Phase examinations, scored using the Backster test data analysis model, can discriminate deception and truth-telling at rates that are greater than chance.

Method

Monte Carlo Design

Monte Carlo[4] methods were used to calculate a dimensional profile of criterion accuracy for the Backster You-Phase technique. The Monte Carlo space consisted of a mathematical simulation of 100 Backster You-Phase examinations. The criterion status of the cases in the Monte Carlo space was determined by comparing a series of random numbers to a fixed base-rate of .5. For each

[3] As described in the presentation of the study data at the annual conference of the American Polygraph Association in Indianapolis (August 2008).

[4] Monte Carlo models are computer intensive statistical methods used to study complex and intangible problems through the use of statistical modeling.

case, the criterion status was set to truthful if the random number was greater than .5. Grand total scores were simulated for the Monte Carlo cases by standardizing another series of random numbers to normative seed parameters for deceptive and truthful cases.

Normative Seeds

Seed parameters for the present Monte Carlo study of Backster were the unweighted average of the three-chart means and standard deviations of the Honts et al. (1985), and Meiron et al. (2008), sampling distributions of deceptive and truthful grand total scores. Although it would be unwise to attempt to assert the representativeness or generalizability of the sampling distributions or results from either the Honts et al. (1985) or the Meiron et al. (2008) studies, the composite of these two sampling distributions can be assumed to contain diagnostic variance, along with error and uncontrolled variance pertaining to Backster numerical scores of examinations conducted with the You-Phase technique.

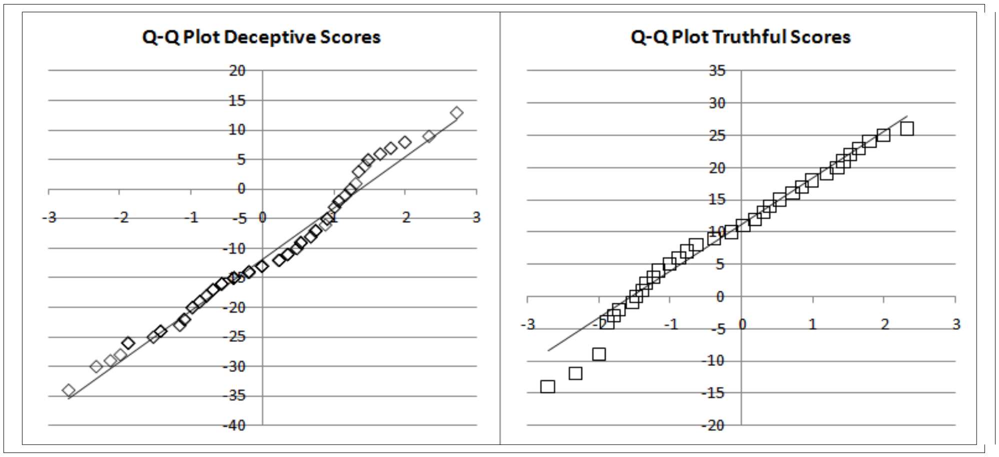
Results

All statistical analyses were completed with a level of significance set at alpha = .05, except as labeled otherwise.

Normative parameters

Before combining the normative parameters from the Honts et al. (1985) and Meiron et al. (2008) samples the data were evaluated for normality. Quantile plots for three-chart deceptive and truthful total scores are shown in Figure 1. Despite the observance of some outlier scores and mild to moderate departures from linearity, it was determined that the data were sufficiently normal to proceed with the construction of the Monte Carlo model with the assumption that total scores of field examinations could be simulated by standardizing random numbers to a standard normal distribution whose parameters are the composite of the distribution parameters from the Honts et al. (1985) and Meiron et al. (2008) samples.

Figure 1. Quantile plots for three-chart Backster numerical scores of You-Phase examinations.



Total scores from the Honts et al. (1985) and Meiron et al. (2008) studies (shown in Table 2), although both considered inadequate as generalizable representations of the population distribution, were compared to each other using an unbalanced two-by-two ANOVA (sample x case status), using the harmonic mean of the sample sizes. Unbalanced ANOVA, using the harmonic mean of the sample sizes, was necessary due to differences in sample sizes. The mean deceptive three-chart scores from the Meiron et al. (2008) study was -12.420 (SD = 8.911), while the mean three-chart truthful score was

10.640 (SD = 7.287). The Honts et al. (1985) study produced a mean deceptive score of -12.500 (SD = 7.794) and a mean truthful score of 3 (SD = 13.856).[5]

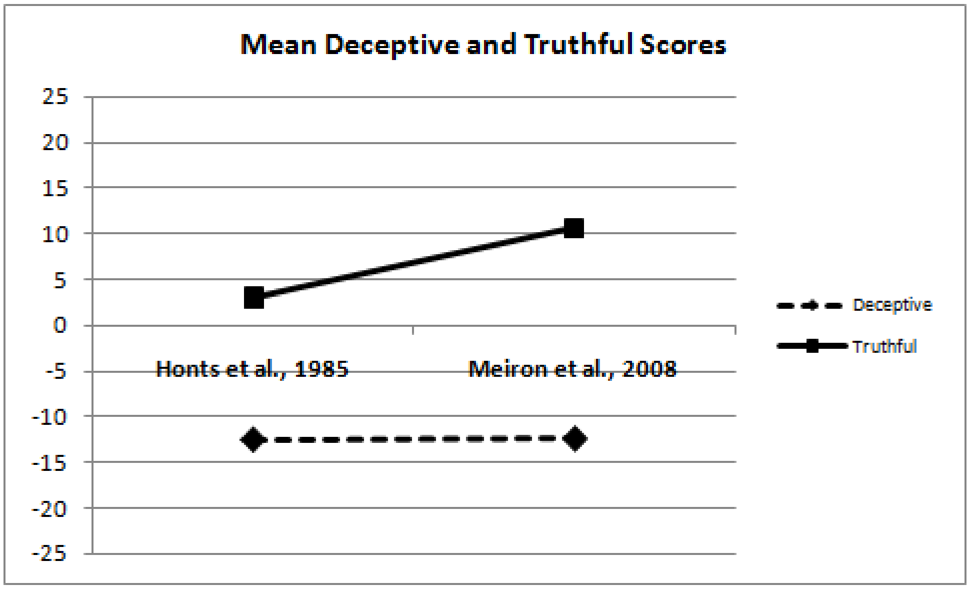
The interaction of sample and case status was significant [F (1,120) = 63.995, (p < .001)]. The difference in the pattern of scores (shown in Figure 2) for the two study samples prevented interpretation of the main effects from the two-way analysis. Post hoc one-way ANOVAs shows that the difference in sampling means of deceptive scores was not significant [F (1,37) = 0.000, (p = .975)], while the difference in sampling means for truthful cases was approaching a significant level [(1,37) = 3.421, (p = .072)]. The unweighted means of the sample distribution parameters were as follows: deceptive mean = -12.460 (SD = 8.353), truthful mean = 6.820 (SD = 10.572). These statistics were the seed parameters for the Monte Carlo model. Because manual test data analysis of PDD examinations is conducted with integer scores, all numerical scores and cutscores in the Monte Carlo model were rounded to the nearest integer.

Table 2. Dimensional profile of criterion accuracy.

	2-chart cutscores +5/-9	Alpha = .05/.05 +1/-11	Alpha = .01/.01 +7/-19	Alpha = .10/.10 -2/-7
Unweighted Average Accuracy	.927 (.032) {.859 to .985}	.918 (.033) {.845 to .976}	.967 (.036) {.878 to .999}	.871 (.036) {.800 to .937}
Unweighted Inconclusives	.313 (.066) {.186 to .442}	.367 (.071) {.231 to .5}	.782 (.057) {.667 to .884}	.128 (.048) {.041 to .227}
Sensitivity	.668 (.068) {.533 to .8}	.575 (.071) {.435 to .712}	.208 (.056) {.103 to .327}	.756 (.062) {.623 to .873}
Specificity	.592 (.071) {.453 to .729}	.724 (.064) {.600 to .843}	.52 (.072) {.386 to .667}	.803 (.058) {.685 to .913}
D INC	.313 (.066) {.186 to .442}	.220 (.058) {.107 to .333}	.470 (.072) {.326 to .605}	.084 (.040) {.019 to .167}
T INC	.329 (.067) {.200 to .465}	.367 (.071) {.231 to .500}	.782 (.057) {.667 to .884}	.128 (.048) {.041 to .227}
FN	.019 (.020) {.001 to .065}	.056 (.032) {.001 to .122}	.010 (.014) {.001 to .043}	.113 (.046) {.033 to .217}
FP	.079 (.039) {.019 to .160}	.058 (.034) {.001 to .137}	.010 (.014) {.001 to .043}	.117 (.044) {.040 to .208}
PPV	.894 (.052) {.784 to .975}	.913 (.050) {.811 to .999}	.953 (.067) {.778 to .999}	.868 (.053) {.759 to .958}
NPV	.969 (.032) {.897 to .999}	.924 (.044) {.829 to .999}	.981 (.026) {.917 to .999}	.874 (.049) {.775 to .958}
D Correct	.972 (.029) {.903 to .999}	.908 (.052) {.788 to .999}	.953 (.066) {.8 to .999}	.866 (.051) {.758 to .955}
T Correct	.882 (.056) {.765 to .974}	.928 (.041) {.844 to .999}	.981 (.025) {.917 to .999}	.876 (.050) {.771 to .962}

[5] These parameters were calculated from the mean and standard errors shown in Figure 1 of Honts, Hodes & Raskin (1985) as measured to the nearest 1/2 point.

Figure 2. Mean deceptive and truthful scores from Honts et al. (1985) and Meiron et al. (2008).



Criterion accuracy

The Monte Carlo space was recalculated for 10,000 iterations. A dimensional profile of criterion accuracy was calculated, including means, standard errors, 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. Cutscores for Backster numerical scores of You-Phase examinations are not based on published studies or normative data. The Backster School of Lie Detection (2011) recommends that three-chart grand total scores of +7 or greater should result in truthful classifications, and that three-chart grand total scores of -13 or lower warrant a deceptive classification. Subtotal scores are not used with the Backster model for test data analysis. Unweighted decision accuracy, using the recommended three-chart cutscores was .956, along with an unweighted inconclusive rate of .512. In addition to the high rate of decision accuracy and high inconclusive results, test sensitivity to deception was worse than chance at .478, with a test specificity level of .520.

Because the Backster School of Lie Detection (2011) allows for a conclusion based on two of three charts, results were calculated using the cutscores recommended for decisions based on two charts. The two-chart grand total cutscore for deceptive classifications is -9, and the two-chart grand total cutscore for truthful classifications is +5. These cutscores correspond to alpha levels (p-values) of .073 and .017 for deceptive and truthful classifications, using the composite norms from Meiron et al. (2008) and Honts et al. (1985) studies. The results are shown in Table 2. Unweighted decision accuracy using the two-chart cutscores was statistically significantly greater than chance (p < .001), along with significantly greater than chance test sensitivity to deception (p < .05). Also shown in Table 2 are the results using cutscores with alpha levels set at .05, .01, and .10.

Discussion

Results from this Monte Carlo study support the validity of the hypothesis that numerical scores of confirmed You-Phase examinations scored with the Backster numerical scoring system can differentiate deception from truth-telling at rates that are significantly greater than chance.

It is clear from the results of this study that the presently recommended cutscores may be suboptimal for decisions based on three charts. These data suggest that attempts to standardize procedures using the recommended cutscores for three-charts will lead to high rates of inconclusive results. Cutscores selected to correspond to alpha levels of .1 for both deceptive and truthful classifications (-2/-7) produced an acceptably high level of decision accuracy and a tolerably low rate of inconclusive results. All other cutscores, including those for two and three chart decisions, produced excessive rates of inconclusive results. Inconclusive results were loaded on truthful cases. Future research on Backster You-Phase examinations should investigate the use of normative data to develop statistically optimal cutscores that can make use of all obtained data and more effectively prioritize objectives regarding test sensitivity, specificity, error rates, and inconclusive results. Decision rules, involving the potential use of both grand total and subtotal scores, should also be the focus of future research.

The most obvious limitation of the present study involves the study design as a Monte Carlo simulation based on sub-optimal seed data from two studies that are themselves unsuitable as criterion studies. Although obviously suboptimal, seed parameters for this Monte Carlo study are based on both field and laboratory cases. Regardless of the acknowledged limitations of the previous studies on which seed parameters are based, the composite distribution parameters can be assumed to be composed of diagnostic, error and uncontrolled variance pertaining to Backster numerical scores of You-Phase examinations. While exact proportions of the different components of variance will remain unknown, the results of this study are encouraging.

Monte Carlo models are considered slightly optimistic, and the present results will be regarded by some as overestimating the criterion accuracy that can be achieved by Backster numerical scores of You-Phase exams. Alternatively, proponents and adherents of the Backster techniques may be inclined to argue that the present results are an underestimation of criterion accuracy. These speculations are ultimately a matter for future research, and the substitution of opinion for evidence is neither warranted nor responsible. Given that field examiners are motivated to achieve results and employ field practice procedures, such as conducting additional test charts or repeating entire examinations, it is likely that the present results are an overestimation of the rate of inconclusive results that would be observed in field settings.

Monte Carlo studies are not regarded, and not intended, as exemplary of the final answer regarding questions of scientific study. Instead, Monte Carlo models and Monte Carlo studies are useful to gain information and insight into complex and intangible problems when other methods of study are not available. Monte Carlo results should not be used in isolation from other studies. Comparison of the present study results with the results of future laboratory and field studies of the Backster You-Phase technique is recommended.

References

Amsel, T. T. (1999). Exclusive or nonexclusive comparison questions: A comparative field study. *Polygraph*, 28, 273-283.

Backster School of Lie Detection (2011). *Basic polygraph examiner's course chart interpretation notebook*. Backster School of Lie Detection: San Diego.

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

Backster, C. (2001). A response to Krapohl & Ryan's 'belated look at symptomatic questions'. *Polygraph*, 30, 213-215.

Bell, B. G., Raskin, D. C., Honts, C. R. & Kircher, J.C. (1999). The Utah numerical scoring system. *Polygraph*, 28(1), 1-9.

Capps, M. H. (1991). Predictive value of the sacrifice relevant. *Polygraph*, 20, 1-6.

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

Gordon, N. J., Mohamed, F. B., Faro, S. H., Platek, S. M., Ahmad, H. & Williams, J.M. (2005). Integrated zone comparison polygraph technique accuracy with scoring algorithms. *Physiology & Behavior*, 87(2), 251-254.

Honts, C., Amato, S. & Gordon, A. (2000). Validity of outside-issue questions in the control question test. Boise State University. Final Report on Grant no. N00014-98-1-0725. DTIC AD Number A376666.

Honts, C., Amato, S. & Gordon, A. (2004). Effects of outside issues on the comparison question test. *Journal of General Psychology*, 131(1), 53-74.

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.

Horvath, F.S. (1988). The utility of control questions and the effects of two control question types in field polygraph techniques. *Journal of Police Science and Administration*, 16, 198-209.

Horvath, F.S. (1991). The utility of control questions and the effects of two control question types in field polygraph techniques. *Polygraph*, 20, 7-25.

Horvath, F.S. (1994). The value and effectiveness of the sacrifice relevant question: An empirical assessment. *Polygraph*, 23, 261-279.

Horvath, F. & Palmatier, J. (2008). Effect of two types of control questions and two question formats on the outcomes of polygraph examinations. *Journal of Forensic Sciences*, 53(4), 1-11.

Kircher, J.C., Kristjiansson, S.D., Gardner, M.K. & Webb, A. (2005). Human and computer decision-making in the psychophysiological detection of deception. University of Utah.

Krapohl, D.J. & Ryan, A.H. (2001). A belated look at symptomatic questions. *Polygraph*, 30, 206-212.

Matte, J.A. (1978). Polygraph Quadri-Zone Comparison Technique. *Polygraph*, 7(4), 266-280.

Matte, J.A. (1996). *Forensic Psychophysiology Using the Polygraph*. Williamsville, NY: J.A.M. Publications: New York.

Matte, J. A. (2001). Reply to rejoinder by Donald J. Krapohl and Andrew H. Ryan. *Polygraph*, 30, 220-222.

Matte, J. A. & Reuss, R.M. (1989). A field validation study of the Quadri-Zone Comparison Technique. *Polygraph*, 18, 187-202.

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

Palmatier, J.J. (1991). Analysis of two variations of control question polygraph testing utilizing exclusive and nonexclusive controls. Unpublished doctoral dissertation:

Podlesny, J.A. & Raskin, D.C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology*, 15, 344-359.

Raskin, D.C. & Hare, R.D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology*, 15, 126-136.

Raskin, D., Kircher, J.C., Honts, C.R. & Horowitz, S.W. (1988). A study of the validity of polygraph examinations in criminal investigations. Final Report, National Institute of Justice, Grant No. 85-IJ-CX-0040.

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

Weaver, R.S. (1980). The numerical evaluation of polygraph charts: Evolution and comparison of three major systems. *Polygraph*, 9, 94-108.

Appendix A

Backster Chart Analysis Rules

Primary Rules

1. Either-Or Rule
2. Non-reinforcement Rule
3. Green Zone 'Yes' Answer Penalty Rule
4. Change in Amplifier Sensitivity Rule
5. Timely Reaction Rule
6. Anticipatory Reaction Rule
7. Lack of Reaction Via Deduction Rule
8. Delayed Cardio Reaction Recovery Rule
9. Minimum Lack-of Reaction Rule
10. Plunging GSR Baseline Rule

Secondary Rules

1. Green Zone Abuse Rule
2. Tracing Average Trend Change Rule
3. Presence of Reaction Via Deduction Rule
4. Single Cycle Trend Conformance Rule

Upgrading Rules

1. Question Pacing Upgrading Rule
2. Tracing Purity Upgrading Rule
3. Reaction Intensity Upgrading Rule

Tracing Oddity Rules

1. Listening Reaction vs. Listening Distortion Rule
2. Answer Reaction vs. Answer Distortion Rule
3. Stabilized Blood Pressure Trend Rule
4. Extra-Systole Cluster Rule