Replication: Criterion Validity of the Empirical Scoring System with Inexperienced Scorers

Raymond Nelson, Mark Handler, and Stuart Senter[1]

Abstract

The authors replicated and extended the work of earlier studies on the Empirical Scoring System (ESS) by calculating criterion accuracy profiles for the ESS and two other scoring systems: 7-position and 3-position. ESS results were also compared to an unweighted ESS model. Data were obtained from a cohort of inexperienced polygraph examiner trainees who evaluated a sample of confirmed single-issue three-question zone comparison test (ZCT) exams (N = 60) from a confirmed case archive constructed by the Department of Defense during 2002. Results are provided from a step-wise ROC analysis of the nonparametric transformation of visibly perceptible response magnitude differences to numerical scores using the bigger-is-better rule. Inexperienced scorers produced results that were equivalent to those of previous studies with 86% decision accuracy (95% CI = 76% to 95%). Ten percent (10%) of the results were inconclusive (95% CI = 3% to 17%). There were no statistically significant differences in decision accuracy or errors for the ESS, 7-position and 3-position systems. However, results with the ESS tended to produce significantly fewer inconclusive results with significant increases in test sensitivity to deception. There were no significant differences in false-positive or false-negative errors for the ESS, 7-position, 3-position, or unweighted ESS models. Results from this study provide additional support for the validity of the ESS. Continued interest in the ESS is recommended.

"Divide each difficulty into as many parts as is feasible and necessary to resolve it."
Rene Descartes

The Empirical Scoring System (ESS; Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson, & Hicks, 2010; Krapohl, 2010; Nelson et al., 2011: Nelson & Krapohl, 2011; Nelson, Krapohl & Handler, 2008) was developed with the goal of establishing an evidence-based method for manually scoring comparison question tests (CQT) in psychophysiological detection of detection (PDD) settings. The ESS was designed around the simplest available solutions that anchor the procedural and empirical validity of the CQT. It is premised on a requirement for published evidence of validation for all procedures and assumptions that define the ESS, including physiological features, numerical transformations, decision rules, and cutscores.

The present study is a replication and extension of earlier studies on the ESS, using a cohort of inexperienced participants who provided blind scores using a sample of confirmed field exams from criminal investigations. Analyses include the comparison of criterion validity, errors, and decision agreement between the ESS and other scoring systems, including an unweighted version of the ESS, 7-position and 3-position scoring systems.

[1] The views expressed in this article are those of the authors, and do not necessarily represent the opinions or policies of Lafayette Instrument Company, the US Government or the Department of Defense. The authors are grateful to Chuck Slupski of the American International Institute of Polygraph and the seven trainees who participated in this study.

Method

Participants

A cohort of seven inexperienced polygraph examiners participated in this study. Participants were trainees in their fifth week of instruction at a polygraph school accredited by the American Polygraph Association. Participation in the study was not mandatory, and had no effect on the employment, performance grades, or training status of the participants. Five of the seven participants were experienced law enforcement investigators with an average of more than 10 years in policing. The remaining two participants had completed their undergraduate degrees in social and behavioral sciences. Six of the participants were male, one female. There were no trainees in the cohort who did not participate in the study and age data for the cohort were not collected.

Sample Data

Sample data were confirmed PDD examinations (N = 60) obtained from the confirmed case archive at the National Center for Credibility Assessment. All sample examinations were single-issue criminal investigation exams conducted using the Federal Zone Comparison Technique (ZCT) question sequence. All cases consisted of three relevant questions, three probable-lie comparison questions and three test charts. Thirty cases were confirmed as truthful and thirty matching cases were confirmed as deceptive. Cases were randomly assigned to six subsets of ten cases each, with no requirement for an equal number of deceptive and truthful cases within the six subsets.

Design and Analysis

Each of the seven participants was randomly assigned two different subsets of 10 cases to score. Study participants scored the first subset using the 7-position system after eight hours of instruction on the procedures described by the Department of Defense Polygraph Institute (2006). No feedback was provided to the participant regarding performance during or after the 7-position scoring task. Participants scored a second subset of different cases the following day after receiving two hours of instruction in using the ESS. Each examination was scored by at least one participant during each of the two scoring conditions. Because six subsets of cases were randomly assigned to the seven participants, two of the subsets were scored by two participants during each of the two scoring conditions. None of the examinations were scored more than once by any of the participants and no participant scored the same case using both scoring methods.

A total of 140 scored results were obtained from the study participants: 70 scores using the 7-position scoring system, and 70 using the ESS. ESS scores were later reduced to their unweighted equivalents by dividing the electrodermal activity scores in half. Seven-position scores were also reduced to their 3-position counterparts. Monte Carlo models were later seeded with the sub-total and total scores.

Empirical Scoring System

A complete description of the ESS procedures was provided by Nelson et al., (2011). A brief description of the procedures can be found in Appendix A, and normative reference data for ESS scores of ZCT examinations can be found in Appendix B. Among the most important principles central to the ESS is the *bigger-is-better* rule, used to assign scores based on visibly observable differences when evaluating responses to relevant and comparison stimuli.

Bigger is Better Rule. We were unable to locate any previously published study of the efficacy of the *bigger-is-better* rule. To investigate the validity of this rule, a step-wise analysis was conducted of the Area Under the Curve (AUC) using the Receiver Operating Characteristic (ROC) statistic (Swets, 1996).[2] Data were expressed in mathematical ratios of the linear measurements of the physiological

[2] AUC can be understood as the ratio or proportion of true positives to false positives when calculated across all possible cutscores. A test with higher sensitivity and higher specificity will produce a greater AUC, and a theoretically perfect model will produce an AUC of 1.

data, using automated measurements taken from an archival sample (N = 292) that was previously used by Nelson et al. (2008). Peak AUC was calculated using progressively smaller response magnitude differences.

Ratios were reduced progressively in a step-wise manner, independently for each component, and for the weighted composite of all components, to locate the peak AUC at which the ratio of true positive and false positive results was most efficient. Table 1 shows the results of the step-wise analysis. Table 2 shows the peak AUC values, and Figure 1 shows the ROC curves for the components and weighted composite.

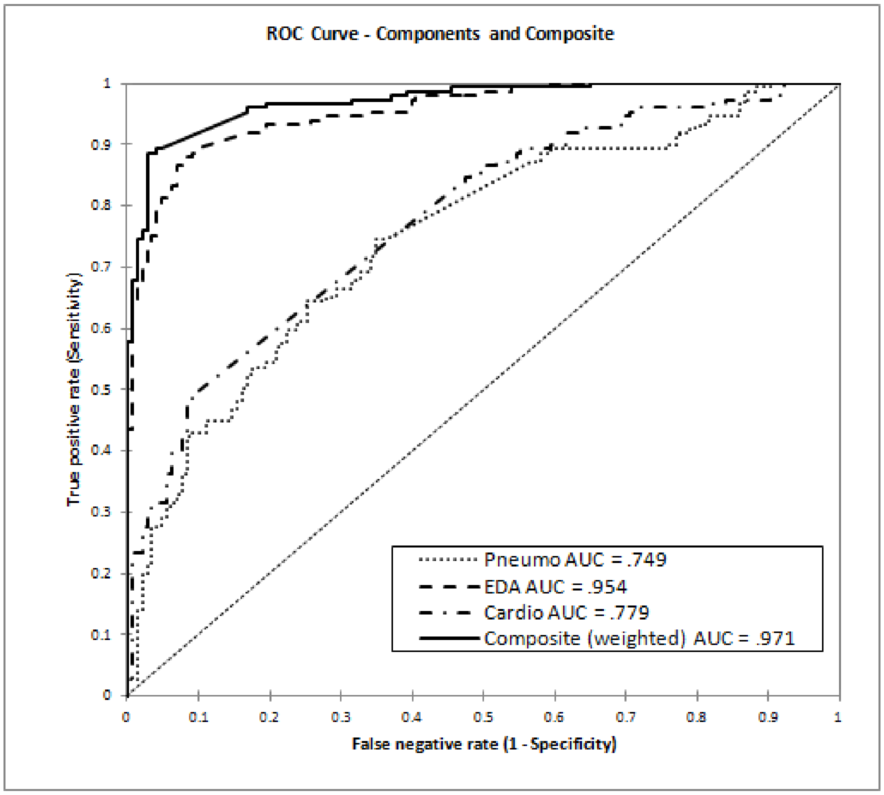
Table 1. Step-wise ROC results. AUC (standard error) and {95% confidence interval}

	Pnuemo	EDA	Cardio	Uniform
1.15:1	.652 (.032) {.589 to .714}	.945 (.012) {.922 to .968}	.748 (.028) {.693 to .803}	.936 (.013) {.912 to .961}
1.14:1	.657 (.032) {.595 to .719}	.947 (.012) {.924 to .970}	.75 (.028) {.695 to .804}	.941 (.012) {.917 to .965}
1.13:1	.659 (.031) {.598 to .721}	.948 (.012) {.925 to .970}	.758 (.028) {.704 to .812}	.944 (.012) {.920 to .967}
1.12:1	.678 (.031) {.617 to .739}	.948 (.012) {.926 to .971}	.767 (.027) {.713 to .820}	.949 (.011) {.928 to .971}
1.11:1	.686 (.031) {.626 to .747}	.948 (.012) {.926 to .971}	.772 (.027) {.72 to .825}	.954 (.011) {.933 to .974}
1.10:1	.702 (.03) {.643 to .761}	.951 (.011) {.929 to .973}	***.779 (.027)*** ***{.727 to .831}***	.958 (.010) {.938 to .978}
1.09:1	.698 (.030) {.639 to .757}	.952 (.011) {.930 to .974}	.774 (.027) {.722 to .827}	.962 (.009) {.944 to .981}
1.08:1	.690 (.031) {.630 to .750}	.953 (.011) {.931 to .974}	.774 (.027) {.722 to .827}	.966 (.009) {.949 to .984}
1.07:1	.705 (.03) {.646 to .764}	.953 (.011) {.931 to .975}	.771 (.027) {.719 to .824}	.968 (.009) {.950 to .985}
1.06:1	.713 (.03) {.654 to .771}	.953 (.011) {.931 to .975}	.767 (.027) {.714 to .821}	.970 (.009) {.953 to .987}
1.05:1	.715 (.030) {.657 to .773}	.953 (.011) {.932 to .975}	.767 (.027) {.714 to .821}	***.971 (.009)*** ***{.954 to .987}***
1.04:1	.738 (.029) {.681 to .795}	***.954 (.011)*** ***{.932 to .975}***	.772 (.027) {.720 to .825}	.971 (.009) {.954 to .987}
1.03:1	.746 (.029) {.69 to .802}	.954 (.011) {.932 to .975}	.771 (.027) {.718 to .824}	.971 (.008) {.954 to .988}
1.02:1	***.749 (.028)*** ***{.694 to .805}***	.954 (.011) {.932 to .976}	.772 (.027) {.720 to .825}	.971 (.008) {.955 to .988}
1.01:1	.746 (.029) {.690 to .802}	.954 (.011) {.932 to .976}	.772 (.027) {.719 to .825}	.971 (.009) {.954 to .987}

Table 2. Maximum areas under the curve and optimal ratios.

	Maximum AUC (95% CI)	Ratio
Pneumograph	.749 (.694 to .805)	1.02:1
Electrodermal	.954 (.932 to .975)	1.04:1
Cardiograph	.779 (.727 to .831)	1.1:1
Uniform (composite)	.971 (.954 to .987)	1.05:1

Figure 1. Maximum areas under the curve with optimal ratios.



These data show the bigger-is-better rule to be a reasonable and valid principle with which to assign numerical scores to PDD responses. This finding should not be surprising, considering the demonstrated effectiveness of computer scoring algorithms (Honts & Driscoll 1987, 1988; Kircher & Raskin, 1988; MacLaren & Krapohl, 2003; Nelson, et al., 2008; Raskin et al., 1988), which are increasingly capable of equaling or exceeding the performance of many human scorers while making use of any measurable difference in response magnitude (Nelson et al, 2008).

Although traditional assumptions and hypotheses regarding the interpretability of linear response ratios are yet incompletely studied, these results support the assumption that physiological reactions of larger magnitude are generally associated with stimuli of greater saliency compared to stimuli that evoke smaller physiological responses.

Weighted and Unweighted ESS

Previous studies have suggested that that EDA contributes approximately one-half of the diagnostic information in polygraph testing (Capps & Ansley, 1992; Harris et al., 2000; Harris & Olsen, 1994; Kircher, Kristjansson, Gardner & Webb, 2005; Kircher & Raskin, 1988, 2002; Krapohl & McManus 1999; Nelson et al., 2008; Raskin, et al., 1988). Therefore, the ESS places 50% of the weight on the EDA data when compositing the pneumograph, electrodermal and cardiograph scores. This is accomplished by doubling all EDA scores, regardless of the magnitude of difference in the strength of reaction, and then summing the scores.

To evaluate the effect of doubling EDA values in the ESS scoring condition, EDA scores were decremented from 7-position to 3-position scores that were weighted equally with scores from the other component sensors. Cutscores were obtained for the unweighted EDA condition using Monte Carlo analysis of the ESS scores reported in Nelson et al. (2008).

7-Position Scoring System

Seven-position scores in the present study were based on the procedures described by the Department of Defense Polygraph Institute (2006), which employs 12 reaction features, and a 7-position numerical transformation rubric in which integer scores (i.e., -3, -2, -1, 0, +1, +2, +3) are assigned to each presentation of each test stimulus question. In this way, observed reactions to relevant questions and comparison questions were reduced to a single set of numerical values for each presentation of each relevant question. See Department of Defense Polygraph Institute (2006) for a full description of conventional manual scoring rules.

Decision rules for the 7-position scoring system are a composite of grand total and sub-total decision rules (e.g. the “spot score rule,” Light, 1999) in which a deceptive result can be determined by either the grand total or subtotal, while a truthful result requires minimum scores for both the grand total and every subtotal (“spot score”). Cutscores for the 7-position system were those that have been used traditionally in field practice and past studies on 7-position manual scoring. Traditional ZCT cutscores evolved to their current standard usage before they were scientifically assessed using normative data. However, traditional cutscores have been the subject of considerable research (Blackwell, 1998; Krapohl, 2005; Yankee, Powell & Newland, 1985). Statistical norms data for 7-position scores of ZCT exams were published by the APA (2011).

3-Position Scoring System

7-position scores were decremented to their corresponding 3-position values. 3-position scoring systems have been described repeatedly in published studies (Blackwell, 1998; Capps & Ansley, 1992; Krapohl, 1998; Harwell, 2000; van Herk, 1991) and have been subject to some criticism for contributing to increased inconclusive results when cutscores are not adjusted for differences in the distributions of 7-position and 3-position scores. Krapohl (1998) showed that adjustment of cutscores could reduce the proportion of inconclusive results arising from 3-position scoring systems, and that the simpler system may be capable of performing equivalently with the 7-position system. However, field examiners have largely continued to rely on traditional cutscores intended for the 7-position scores when using the 3-position system (Department of Defense Polygraph Institute, 2006), and 3-position results in this study were evaluated using traditional cutscores.

Analysis

Decision accuracy, errors and inconclusive rates for criterion deception and criterion truthful cases were calculated. Bootstrap resampling was used to estimate the variance of the distributions of values for the weighted and unweighted ESS models, and the 7-position and 3-position scoring systems. Positive predictive values (PPV) and negative predictive values (NPV) were calculated with the assumption of uniform prior probabilities for truthful and deceptive conditions. PPV is the proportion of true-positive results to all positive results, and can be understood as the estimated probability that a positive result or "failed" test is correct.

NPV is the proportion of true-negative results to all negative results, and can be understood as the estimated probability that a negative result or "passed" test is correct. A series of pairwise analyses were conducted, using bootstrap t-tests, when the results of the bootstrap distributions indicated a statistically significant difference in results between the scoring conditions.

Results

Table 3 shows bootstrap mean decision accuracy and inconclusive rates, and 95% confidence intervals for the cohort of inexperienced examiners who scored the sample cases using the 7-position scoring system and the ESS.

Table 3. Mean, (Standard Deviations) and {95% Confidence Intervals} for All Scoring Conditions

	Decision Accuracy	**Inconclusive Results**
Empirical Scoring System	85.7% (4.8) {76.2 to 94.9}	10.1% (3.9) {2.5% to 17.7}
7-Position System	81.5% (5.7) {70.3 to 92.6}	22.8% (7.5) {11.9 to 33.6}
3-Position System	87.2% {76.7 to 97.7}	32.8% {20.9 to 44.5}
Unweighted ESS	85.5% {75.5 to 95.5}	21.4% {11.1 to 31,8}

It was not possible to calculate statistical confidence intervals for NPV for the 7-position and 3-position systems because there were no false-negative errors in the present study. Evaluation of the NPV for the weighted and unweighted ESS models showed that all statistical confidence intervals include the ceiling value of 100%. In other words, no statistically significant difference existed between the NPV of the ESS and other scoring systems in this study.

A double bootstrap of the results of the ESS and 7-position scoring systems showed the mean rate of decision agreement for ESS and 7-positions results was .88 (95% CI = 77% to 99%).

Tables 4-7 show pairwise comparisons between the five decision approaches using the results from 7-position scoring and the ESS. Although most differences were not significant, there was a statically significant reduction in inconclusive results for the ESS (p < .01) compared to the 7-position system (Table 4), that was loaded on criterion deceptive cases. Additionally, there were statistically significant increases in both test sensitivity and test specificity. Differences in false-negative and false-positive errors were not statistically significant.

Table 5 shows a comparison of ESS and 3-position results. Decision accuracy for 3-position system did not differ significantly from the ESS. However, there were significantly fewer inconclusive results in the ESS condition for both truthful and deceptive cases, in addition to statistically significant increases in test sensitivity and test specificity. Differences in false-negative and false-positive errors were not significant.

Table 4. Differences in Decision Outcomes for ESS and 7-Position Scores

	ESS	**7-Position**	**Sig.**
Decision Accuracy	85.7% (4.8) {76.2 to 94.9}	81.5% (5.7) {70.3 to 92.6}	(p = .22)
Inconclusive Results	10.1% (3.9) {2.5% to 17.7}	22.8% (7.5) {11.9 to 33.6}	(p < .01)*
Sensitivity	94.6% {86.9 to 1}	85.8% {73.1 to 98.5)	(p = .05)*
Specificity	57.3% {39.0 to 75.6}	40.0% {22.4 to 57.6}	(p = .03)*
Inconclusive Truthful	18.4% (7.2) {4.3 to 32.5}	31.4% (8.7) {14.4 to 48.4}	(p = .05)*
Inconclusive Deceptive	2.6% (2.9) {0 to 8.3}	14.2% (6.5) {1.5 to 26.8}	(p = .01)*
False-positive Errors	24.3% {8.5 to 40.2}	28.6% {12.5 to 44.7}	(p = .30)
False-negative Errors	2.8% {0 to 8.5}	0.0% {0 to 0}	(p = .09)
Positive Predictive Value	81.2% (6.5) {68.5 to 93.9}	74.9% (8.6) {60.4 to 89.5}	(p = .18)
Negative Predictive Value	94.9% (5.3) {84.5 to 100}	(N/A)	(N/A)

* statistically significant difference

Table 5. Differences in Decision Outcomes for ESS and 3-Position Scores

	ESS	**3-Position**	**Sig.**
Decision Accuracy	85.7% (4.8) {76.2 to 94.9}	87.3% (5.2) {77.0 to 97.5}	(p = .38)
Inconclusive Results	10.1% (3.9) {2.5% to 17.7}	32.7% (6.1) {20.7 to 44.8}	(p < .001)*
Sensitivity	94.6% {86.9 to 1}	82.8% {69.4 to 96.3}	(p = .02)*
Specificity	57.3% {39.0 to 75.6}	35.5% {17.9 to 53.0}	(p = .01)*
Inconclusive Truthful	18.4% (7.2) {4.3 to 32.5}	46.9% (9.2) {28.8 to 65.0}	(p < .001)*
Inconclusive Deceptive	2.6% (2.9) {0 to 8.3}	17.1% (6.9) {3.7 to 30.1}	(p < .01)*
False-positive Errors	24.3% {8.5 to 40.2}	17.6% {3.8 to 31.4)	(p = .20)
False-negative Errors	2.8% {0 to 8.5}	0.0% {0 to 0}	(p = .09)
Positive Predictive Value	81.2% (6.5) {68.5 to 93.9}	82.9% (6.9) {69.5 to 96.3}	(p = .42)
Negative Predictive Value	94.9% (5.3) {84.5 to 100}	(N/A)	(N/A)

* statistically significant difference

Table 6 shows the results of a statistical comparison of 7-position scores and 3-position scores. Differences in test sensitivity and test specificity were not significant, nor were differences in false-negative and false-positive errors. There was a statistically significant difference in the percentage of inconclusive results between the 7-position and 3-position scoring systems. This difference was limited to truthful cases. The 7-position scoring system produced significantly fewer inconclusive results with criterion truthful cases compared to the 3-position system.

Table 6. Differences in Decision Outcomes for 7-Position and 3-Position Scores

	7-Position	**3-Position**	**Sig.**
Decision Accuracy	81.5% (5.7) {70.3 to 92.6}	87.3% (5.2) {77.0 to 97.5}	(p = .14)
Inconclusive Results	22.8% (7.5) {11.9 to 33.6}	32.7% (6.1) {20.7 to 44.8}	(p = .04)*
Sensitivity	85.8% {73.1 to 98.5)	82.8% {69.4 to 96.3}	(p = .32)
Specificity	40.0% {22.4 to 57.6}	35.5% {17.9 to 53.0}	(p = .31)
Inconclusive Truthful	31.4% (8.7) {14.4 to 48.4}	46.9% (9.2) {28.8 to 65.0}	(p = .04)*
Inconclusive Deceptive	14.2% (6.5) {1.5 to 26.8}	17.1% (6.9) {3.7 to 30.1}	(p = .33)
False-positive Errors	28.6% {12.5 to 44.7}	17.6% {3.8 to 31.4)	(p = .08)
False-negative Errors	0.0% {0 to 0}	0.0% {0 to 0}	(N/A)
Positive Predictive Value	74.9% (8.6) {60.4 to 89.5}	82.9% (6.9) {69.5 to 96.3}	(p = .14)
Negative Predictive Value	(N/A)	(N/A)	(N/A)

* statistically significant difference

Table 7 shows the comparison of ESS and unweighted ESS results. Test sensitivity to deception was significantly increased for the weighted ESS model, along with a statistically significant reduction in the proportion of inconclusive results for criterion deceptive cases. Differences in false-negative and false-positive errors were not significant.

Table 7. Differences in Decision Outcomes for ESS with weighted and unweighted EDA

	ESS (Weighted)	Unweighted ESS	Sig.
Decision Accuracy	85.7% (4.8) {76.2 to 94.9}	85.5% (5.1) {75.5 to 95.5}	(p = .50)
Inconclusive Results	10.1% (3.9) {2.5% to 17.7}	21.5% (5.3) {11.1 to 31.8}	(p < .01)*
Sensitivity	94.6% {86.9 to 1}	78.2% {63.9 to 92.5)	(p < .01)*
Specificity	57.3% {39.0 to 75.6}	54.7% {36.3 to 73.1}	(p = .39)
Inconclusive Truthful	18.4% (7.2) {4.3 to 32.5}	24.2% (8.2) {8.2 to 40.3}	(p = .22)
Inconclusive Deceptive	2.6% (2.9) {0 to 8.3}	19.0% (6.9) {5.5 to 32.5}	(p < .001)*
False-positive Errors	24.3% {8.5 to 40.2}	21.1% {5.7 to 36.4}	(p = .34)
False-negative Errors	2.8% {0 to 8.5}	2.8% {0 to 8.5}	(p = .50)
Positive Predictive Value	81.2% (6.5) {68.5 to 93.9}	80.7% (7.1) {66.7 to 94.6}	(p = .47)
Negative Predictive Value	94.9% (5.3) {84.5 to 100}	94.7% (5.6) {83.6 to 100}	(p = .48)

* statistically significant difference

Discussion

The weighted ESS model produced significant increases in test sensitivity to deception and significantly fewer inconclusive results when compared to the 7-position, 3-position and unweighted ESS models. The difference in inconclusive results was primarily loaded on deceptive subjects. Comparison of the 7-position and 3-position systems revealed that the 7-position system also produced significantly fewer inconclusive results among truthful cases, suggesting that weighted scoring models may generally outperform unweighted models.

Although overall differences in decision accuracy were not significant, none of the scoring models in this study outperformed the ESS along any measured dimension of test accuracy. One obvious implication begins to emerge: limiting test data analysis activities to the interpretation of a core set of primary scoring features does not reduce decision accuracy.

The importance of expertise, or lack of expertise among the participants in this study participants, deserves discussion. The goal of applied research is to create, define, and refine knowledge that can be generalized to practical field circumstances. Past studies in PDD scoring accuracy have sometimes relied on expert examiners as participants. Assuming a normal distribution of abilities and skills, 2.5 to 5.0% of professionals can be expected to possess skills that exceed the normal range and deserve the "expert" designation. It is also likely that a somewhat larger proportion of examiners will want to consider themselves experts, but the simple and obvious mathematical fact is that most professionals' skills will be within the normal range. Effectiveness of test data analysis models that are verified as effective with those professionals whose skill level is in the top 2.5 to 5.0% may not be generalizable to the 95.0 to 97.5% of professionals whose skills are in the normal professional range. For this reason, studying PDD decision accuracy under optimal conditions could lead to results that may not be fully generalizable to field conditions.

An optimal-conditions approach to applied research may lack external validity if the model is not also verified as effective with professionals of lesser experience or qualifications working under sub-optimal conditions. A scoring model with known reliability and validity for professionals with skills within the normal or average range of abilities is arguably more applicable and generalizable to the range of individuals and circumstances in field practice settings. Validity and reliability of the PDD test will be most effectively improved by increasing the skills of the majority of examiners rather than a small number of exceptional experts at the extreme end of the skill level continuum. Results from other research involving experienced scorers should be compared with these study results to better understand issues related to expertise in the test data analysis arena.

False negative errors

Some of the statistics of interest could not be calculated from the sample data, including NPV for both 7-position and 3-position systems. The reason for this was an absence of false-negative errors with the sample cases. A second ancillary analysis was completed to help understand the significance and meaning of this observed result. Krapohl (2006) reported an inconclusive rate of 23% among truthful cases for Federal ZCT exams, along with decision accuracy of 82.0%. The false-positive error rate among truthful cases was 14.0% for the studies cited (Blackwell, 1998; Krapohl, 2005; Yankee Powell & Newland, 1985). For deceptive cases, Krapohl (2006) reported an inconclusive rate of 9.0%, decision accuracy at 97.0%, and false-negative errors at a rate of 2.7%.

Using the 2.7% false-negative rate as the expected mean frequency, a Poisson analysis[3] (Haight, 1967) showed that the probability of observing zero false-negative

[3] Poisson analysis can be used to predict the random chance probability of observing a given frequency of rare events (i.e. 0 FN errors) in a time-series sample of a known size (i.e., N = 60) when the expected frequency is known (i.e., 2.7% FN errors reported in previous studies.

errors in a sample of the present size was not significant (p = .44) in a small sample such as that used in this analysis.

Based on these data, it would not be realistic to expect a false negative error rate of zero in field settings. The only reasonable interpretation that can be made regarding the present data is that the confidence intervals for the ESS include the observed of the 7-position and 3-position systems and that no statistically significant difference exists in NPV or FN errors between the ESS and the 7-position and 3-position systems. To explain further, a consequence of the small sample size is a high likelihood of observing zero false negative errors due to random chance alone. Moreover, it is unlikely that any test can be expected to function with perfect accuracy under all conditions.

False-positive errors

The participants in this study produced more false-positive errors than were observed in previous studies on the ESS. This suggests the potential for a scoring bias among the study participants. Future studies should evaluate the role of scoring bias in the results of manual test data analysis paradigms.

Within-test subtotal score differences

A third and final ancillary analysis involved between-subtotal score differences. Within-test differences between subtotal scores are sometimes of interest to field examiners when evaluating data that produce a truthful grand total score while one of the subtotals produces a score that appears inconsistent with the trend of the other subtotals or the total score. The concern to field examiners will be whether or not the subtotal scores provide a reliable and useful opportunity to achieve an accurate deceptive result and reduce the occurrence of false negative errors.

A bootstrap analysis of the distribution of within-test differences between subtotal scores produced a normal distribution of values, with a mean of 3.6 points, and standard deviation of 2.3 points. The mean within-test difference for truthful cases was 4.3 (SD = 2.5) and for deceptive cases was 3.0 (SD = 1.8). Bootstrap calculation of the z-value of the normal distribution of within-test differences in subtotal scores showed the 90th percentile to be 6.7 (95% CI = 5.3 to 8.1) points for all cases, 7.5 (95% CI = 5.4 to 9.8) for truthful cases, and 5.0 for deceptive cases (95% CI = 3.1 to 6.9).

These data suggest that a within-test difference of less than 8 points can be considered to be a normal occurrence among truthful examinees. Differences of 8 or more points between subtotal scores can be reasonably expected to be observed in less than 10% of ZCT exams. Field examiners who desire an evidence-based guideline for the interpretation of extreme within-test differences between subtotals may wish to consider a difference of 8 or more points to be an unusual occurrence, and might be justified in setting aside the grand total result as inconclusive under this condition. However, there were no significant differences in decision accuracy or inconclusive results with either truthful or deceptive cases using both the grand total rule and two-stage rule, when the sub-total scores exceeded 8 points in difference.

Imposition of a more restrictive (i.e., less than 8 points) boundary of subtotal differences resulted in substantial decreases in both test specificity and overall decision accuracy, along with increased inconclusive results for truthful examinees. It appears the use of scientific decision rules, statistically optimal cutscores, and Bonferonni correction will optimize the test accuracy without the need for the additional rule concerning within test differences between subtotal scores. An important consideration is that the addition of rules based on hypothesis alone, in the absence of supporting evidence, may be unwarranted, can be expected to deplete decision accuracy, and might be expected to contribute to test results that lead to unproductive post-test activities in field settings.

Limitations

Despite the encouraging results from this study, several limitations exist. Both the cohort of inexperienced examiners and the confirmed case sample are small in size, meaning that there is some possibility that important conditions that affect fieldwork may not be well represented in the present study or that some statistical results become influenced by the small sample size. Additionally, the present study does not address statistical problems inherent to multi-issue screening exams, and these will need to be addressed in a separate project. The present analysis is not intended to determine the causal relationships for the observed results, and it remains unknown whether the effectiveness of the ESS is due to differences in scoring features, simplified transformations, decision rules, or the statistically derived cutscores. These questions should be addressed in future studies.

One aspect of the study data collection method deserves further discussion, in that all study participants learned and completed ESS scoring tasks after first learning and completing the 7-position scoring task, leaving an unknown potential for a confounding practice effect. A preferable design would have been a counterbalanced analysis in which half of the participants scored the first subset using the ESS while the other half first used 7-position scoring. However, it was not part of the teaching schedule for the small cohort of participants. Instead, each student scored different subsets and no student scored the same subset using both TDA models. However, the potential practice confound is incompletely controlled in this study. Although observed differences were statistically meaningless in this study, replication of this experiment with the opposite sequence of activities or a counterbalance design would help to better understand the role that skill acquisition played in these observed results.[4]

Given the complexity and ambition of this study, involving the comparison of the ESS to 7-position, 3-position, and unweighted ESS models, the use of an ANOVA-based analytic approach might be preferable to some readers. A bootstrapping and frequency-table analysis was thought to be a simpler alternative that also provides potential readers greater information and insight into the effect of scoring condition on various aspects of the test accuracy profile.

Another issue that deserves discussion is the sampling methodology and use of bootstrapping for statistical analysis. There is no great advantage to use of bootstrapping as an analysis method, and the results from bootstrapping are not expected to differ markedly from those of other methods. However, bootstrapping does offer the advantage of the calculation of empirical confidence intervals that do not depend on assumptions about distribution shape. In addition to providing a robust analysis method, use of bootstrap resampling mitigated difficulties related to the number of participants and to group inequality in the sub-samples.

Bootstrapping, though considered robust against some departures in sampling distribution shape, cannot overcome the inherent limitations of a small sample and small cohort of participants. Use of a cohort of inexperienced examiners may partially overcome concerns about overestimation, but this does not overcome the overarching limitation that the results herein are informative only so long as it is assumed that the sample and cohort are representative of the population of examinees and other scorers.

An ideal sample of bootstrapping seed data would have included both randomly selected examinations for which each exam was scored by a different randomly selected examiner, thereby assuring complete independence of each examination. A practical limitation on this ideal is that field samples are never random as they depend on the non-random case confirmation as a selection mechanism.

An alternative design might be to have all examiners score all cases using all scoring systems. However, an exhaustive pair sampling method might be expected to

[4] Generally recognized advantages of within-subjects (repeat measures) designs include the potential for more effective use of available participant resources, including the potential for greater statistical power compared with between-subject designs, and the potential reduction of confounding effects resulting from individual differences.

increase the introduction of both practice and fatigue effects into the sampling scores. Additionally, exhaustive pairing of exam and scorer for all conditions would have decreased the degree of independence among the sampling scores because similarity of scores would be more greatly influenced by the fact that the same scorers would have provided all scores for all scoring conditions using the same sample cases.

Because there was no expectation of complete independence in this study, due to the intent to complete the analysis with replication scores using a single case sample under different scoring conditions, the choice to reduce the scoring task to 10 cases for each participant represents a reasonable compromise that may have both increased the degree of independence among the data in the different scoring conditions (i.e., by reducing the uniform pairing of case and scorer in both conditions), and reduced the impact of practice and fatigue effects.

Another limitation was that this project did not attempt to determine the statistical effect of the major differences between the ESS and other scoring systems.[5] Instead the goal of this study was to provide statistical confidence intervals for criterion accuracy profiles of the ESS and other scoring systems, with the hope that confidence intervals would offer accessible information to both scientific and non-scientific readers. Inquiry into causal and structural differences should become the basis of another study. A final limitation, due to the study design and the nature of the data collection method, was interrater agreement was not calculated for the study data. Past studies on the ESS have shown it to provide interrater agreement that equals or exceeds that of other scoring systems (Blalock et al., 2009; Nelson et al., 2008).

Conclusion

The results of this study provide several insights into manual scoring systems. These data suggest that an empirically developed weighted 3-position scoring system may produce accuracy rates that are equivalent to the 7-position system. Although no single study is sufficient to anchor empirical conclusions, results of this study do provide additional support for the validity of the ESS. Results of this study, together with results of earlier studies, suggest that human scores are capable of using the ESS to achieve high rates of decision accuracy that are equivalent to that of other extant scoring systems that rely on more complex assumptions and procedures.

A practical aspect of a simple empirically based scoring system is that reliable skill acquisition can be expected to increase, and difficulty in skill acquisition can be reduced without loss of test accuracy. Another benefit of a simplified evidence-based scoring rubric is that scoring proficiency becomes a less perishable skill for those examiners whose professional responsibilities include other forms of policing, administration, investigation, research, etc. Moreover, the simpler assumptions of the weighted 3-position transformations of the ESS, compared to unwieldy linear assumptions of 7-position numerical transformations, may provide a more realistic path toward understanding the validity of underlying theoretical constructs regarding the observation and measurement psychophysiological responses to PDD test stimuli.

The ESS appears to provide an attainable and effective system for analyzing the results of PDD examinations and continued research on the ESS is recommended. Future studies should explore the use of normative data and norm-referenced cutscores to better understand differences between various scoring systems. Additionally, increased use of ROC analysis may provide a method that is more resistant to differences in cutscores compared to results based on the statistical hypothesis paradigm. Future studies should continue to compare the results of the ESS with other test data analysis models.

[5] The main differences are 1) a reduced scoring feature set, 2) the use of norm-referenced cutscores, and 3) the use of decision rules derived from scientific studies.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph*, 40(4), 196-305.

Backster, C. (1963). Do the charts speak for themselves? New standards in polygraph chart interpretation. *Law and Order*, 11, 67-68, 71.

Barland, G. H. (1985). A method of estimating the accuracy of individual control question polygraph tests. In, Anti-terrorism; forensic science; psychology in police investigations: Proceedings of IDENTA-'85 (pp. 142-147). The International Congress on Techniques for Criminal Identification.

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

Blackwell, J. N. (1998). PolyScore 33 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations. Available at the Defense Technical Information Center. DTIC AD Number A355504/PAA. Reprinted in *Polygraph*, 28(2), 149-175.

Blalock, B., Cushman, B., & Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. *Polygraph*, 38(4), 281-288.

Capps, M. H., & Ansley, N. (1992). Comparison of two scoring scales. *Polygraph*, 21(1), pp. 39-43.

Department of Defense Polygraph Institute (2006). *Test Data Analysis: DoDPI numerical evaluation scoring system*. Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007.

Haight, F. A. (1967). *Handbook of the Poisson Distribution*. New York: John Wiley & Sons.

Handler, M., Nelson, R, Goodson, W., & Hicks, M. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph*, 39(4), 200-215.

Harris, J., Horner, A., & McQuarrie, D. (2000). An Evaluation of the Criteria Taught by the Department of Defense Polygraph Institute for Interpreting Polygraph Examinations. Johns Hopkins University, Applied Physics Laboratory. SSD-POR-POR-00-7272.

Harwell, E. M. (2000). A comparison of 3- and 7-position scoring scales with field examinations. *Polygraph*, 29(3), 195-197.

Honts, C. R., & Driscoll, L. N. (1987). An evaluation of the reliability and validity of rank order and standard numerical scoring of polygraph charts. *Polygraph*, 16(4), 241-257.

Honts, C. R., & Driscoll, L. N. (1988). A field validity study of rank order scoring system (ROSS) in multiple issue control question tests. *Polygraph*, 17(1), 1-15.

Kircher, J. C., & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, 73(2), 291-302.

Kircher, J. C., & Raskin, D. R. (2002). Computer methods for the psychophysiological detection of deception. In Murray Kleiner (Ed.) *Handbook of Polygraph Testing*. Academic Press: San Diego, CA.

Kircher, J. C., Kristjansson, 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., & McManus, B. (1999). An objective method for manually scoring polygraph data. *Polygraph*, 28(3), 209-222.

Krapohl, D. J. (1998). A comparison of 3- and 7- position scoring scales with laboratory data. *Polygraph*, 27(4), 210-218.

Krapohl, D. J. (2005). Polygraph decision rules for evidentiary and paired testing (Marin Protocol) applications. *Polygraph*, 34(3), 184-192.

Krapohl, D. J. (2006). Validated polygraph techniques. *Polygraph*, 35(3), 149-155.

Krapohl, D. J. (2010). Short Report: A test of the ESS with two-question field cases. *Polygraph*, 39(2), 124-126.

Light, G. D. (1999). Numerical evaluation of the Army zone comparison test. *Polygraph*, 28(1), 37-45.

MacLaren, V., & Krapohl, D. J. (2003). Objective assessment of comparison question polygraph. *Polygraph*, 33(2), 107-126.

Marin, J. (2000). He said/ She said: Polygraph evidence in court. *Polygraph*, 29(4), 299-30.

Marin, J. (2001). The ASTM exclusionary standard and the APA 'litigation certificate' program. *Polygraph*, 30(4), 288-293.

Nelson, R., Krapohl, D. J. & 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(3), 185-215.

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

Raskin, D. C., 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.

Senter, S. M. (2003). Modified general question test decision rule exploration. *Polygraph*, 32 (4), 251-263.

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

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnosis*. Mahwah, NJ: Erlbaum.

Timm, H. W. (1982). Analyzing deception from respiration patterns. *Journal of Police Science and Administration*, 10, 47-51.

Van Herk, M. (1990). Numerical evaluation: Seven point scale +/-6 and possible alternatives: A discussion. *The Newsletter of the Canadian Association of Police Polygraphists*, 7, 28-47. Reprinted in *Polygraph*, 20(2), 70-79.

Yankee, W. J., Powell, J. M, III & Newland, R. (1985). An investigation of the accuracy and consistency of polygraph chart interpretation by inexperienced and experienced examiners. *Polygraph*, 14, 108-117.

Appendix A. Empirical Scoring System

I. Physiological Signals

1. Respiration
   a) Decrease in respiration amplitude for three or more respiratory cycles, beginning after the stimulus onset
   b) Slowing of respiration rate for three or more respiratory cycles, beginning after the stimulus onset
   c) Temporary increase in respiratory baseline for three or more respiratory cycles, beginning after the stimulus onset
2. Electrodermal response amplitude
3. Cardiograph baseline increase, observed at the diastolic baseline

II. Transformations

1. Assign values of +, - or 0 using the 3-position scale and the bigger-is-better rule
   a) Score any visibly discernible difference in response magnitude (without electronic or mechanical measurement)
   b) Do not be concerned about traditional scoring ratios
   c) Score each RQ to the stronger of bracketing comparison questions, for each component sensor
   d) Double all EDA values to +/- 2
2. Score only timely reactions
   a) Do not score reactions that begin before the stimulus
   b) Do not score reactions that begin more than 5 seconds after a timely answer
3. Score only normal interpretable data
   a) Do not attempt to score data that are affected by movement artifacts
   b) Do not attempt to score messy or unstable segments of data
   c) Do not attempt to score data of unusual response quality (dampened or exaggerated)

III. Decision Rules

1. Two-stage decision rules (Senter Rules)
   a) Optimal for most purposes
   b) Increased sensitivity (without excessive increase in FP errors)
   c) Reduced inconclusive results
   d) Use Bonferonni corrected alpha for spot scores
2. Grand Total Rule - highest level of decision accuracy
3. Spot scoring rules
   a) Only for multiple issue screening exams

IV. Normative Data (cutscores)

1. ZCT (single-issue)
   a) Total score >= +2 = alpha <= .10 (NSR/NDI)
      - Total scores >= +5 = alpha <= .05 (NSR/NDI)
   b) Total score <= -4 = alpha <= .05 (SR/DI)
   c) *Any* sub-total <= -7 = Bonferonni corrected alpha <= .0167 x 3 RQs = .05 (SR/DI)
2. MGQT/DLST (multi-issue screening)
   a) *Any* sub-total <= -3 = alpha <= .05 (SR)
   b) *All* sub-totals >= +1 = alpha <= .10 (NSR)
      - *All* sub-totals >= +2 = alpha <= .05 (NSR)

Appendix B. Empirical Scoring System – Normative Reference Data for Single-Issue ZCT Exams with Three Relevant Questions (Nelson, Krapohl, & Handler, 2008)

Mean deceptive score = -9.14 (SD = 8.74)
Mean truthful score = 8.35 (SD = 7.89)

Truthful (NSR) Cutscores

Total NSR Cutscore	p-value (alpha)
1	.106
2	.085
3	.067
4	.052
5	.040
6	.030
7	.023
8	.017
9	.012
10	.008
11	.006
12	.004
13	.003
14	.002
15	.001
16	<.001

Deceptive (SR) Cutscores

Total SR Cutscore	p-value (alpha)
0	.127
-1	.099
-2	.077
-3	.058
-4	.043
-5	.032
-6	.023
-7	.016
-8	.011
-9	.008
-10	.005
-11	.003
-12	.002
-13	.001
-14	<.001

Cutscores
NDI >= +2 (a = .10)
NDI >= +5 (a = .05)
DI <= -4 (p = .05)
Spot scores for DI <= -7 (Bonferroni corrected alpha = .0167)

Appendix C. Two-Stage Decision Rules (Senter 2003; Senter & Dollins, 2008)

Two-stage decision rules provide;

1. Optimal decision rules for most purposes.
2. Balanced sensitivity and specificity to deception.
3. Increased sensitivity to and decreased inconclusive results compared to grand-total decision rule.
4. Protection from excessive increases in false-positive errors that may result when subtotal scores are permitted to supersede the importance of grand total scores.

2-Stage Procedure

Stage 1: Grand Total Only (do NOT use the Spot Score Rule at Stage 1)

A. If the Grand Total >= +2 then NDI (a = .10) or +5 (a = .05)

B. If the Grand Total <= -4 then DI (a = .05)

Stage 2: sub-total Scores (*only* if the Grand Total is inconclusive during Stage 1)

A. If any sub-total (RQ spot) <= -7 then DI (a = .017, Bonferonni correction)

B. There are no NDI considerations using sub-totals at Stage 2

* Decisions based on the sub-total/spot scores, in event-specific/single-issue exams, are made by comparing the sub-total score to the normative distribution of total (not spot) scores.