Human and Computer Decision-Making in the Psychophysiological Detection of Deception

John C. Kircher, Sean D. Kristjansson, Michael K. Gardner, and Andrea Webb[1]

Executive Summary

This is the final report of a project with three objectives. Swinford (1999) described the physiological changes considered to be diagnostic of deception by the Department of Defense Polygraph Institute[2] (DoDPI). The first objective was to assess the reliability and validity of those physiological criteria. Twelve of 25 criteria were diagnostic of deception in a sample of confirmed field cases. However, three had low reliability, one was an artifact of question position, another three did not generalize to an independent sample of laboratory cases, and two did not generalize to an independent sample of field cases. In the end, three of the 25 criteria were valid and reliable indicators of deception across three independent samples of cases and were consistent with published DoDPI criteria.

The second objective was to determine which of the criteria federal polygraph examiners use to evaluate polygraph charts from probable-lie examinations. Lens model analyses of electrodermal, cardiovascular, and respiration data were conducted for each of 32 experienced federal polygraph examiners. The lens models revealed that examiners based their numerical evaluations of electrodermal responses almost exclusively on response amplitude. Numerical evaluations of the cardiograph were based primarily on increase in baseline. 13% of the examiners also used decrease in baseline, and less than 10% used changes in pulse amplitude or pulse rate. Numerical evaluations of respiration were predicted by decrease in line length for 88% of the examiners. Numerical scores for 16% of the examiners were related to increase in respiration baseline. Numerical scores for less than 10% of the examiners were related to decrease in respiration amplitude or decrease in rate.

The third objective was to propose a method for combining the DoDPI criteria in order to maximize the accuracy of polygraph decisions. Computer measures of DoDPI criteria were obtained from a standardization sample of 80 confirmed single-issue field cases and were used to develop three discriminant functions. The three functions were tested on an independent sample of 80 cases from the database of confirmed polygraph examinations maintained by DoDPI. The discriminant function with the greatest accuracy was composed of electrodermal response amplitude,

[1] All authors were with the Department of Educational Psychology at the University of Utah. The original report is available from the Defense Technical Information Center (http://www.dtic.mil/dtic/). Reference number ADA520590.

Acknowledgements
The authors are deeply indebted to the 32 polygraph interpreters who volunteered several days of their time to numerically evaluate the physiological recordings. We also are grateful to the Department of Defense Polygraph Institute (DoDPI) research staff who provided the physiological data, recruited all of the interpreters, and scheduled all of the meetings with the interpreters. We wish to thank the various federal agencies that provided release time to their examiners and gave them the opportunity to participate. We thank the reviewers for many useful comments that improved the final report. Finally, we wish to thank Professor Emeritus David C. Raskin for his review and helpful suggestions on an earlier draft of this report. The findings are in general agreement with Dr. Raskin's research and recommendations he made to the polygraph community more than 25 years ago.

[2] The Department of Defense Polygraph Institute is now the National Center for Credibility Assessment. The original title has been retained throughout this report.

cardiograph increase in baseline, and respiration decrease in line length. Excluding 26.5% inconclusive outcomes, computer decisions were 81.5% correct when the suspect was truthful and 87.9% correct when the suspect was deceptive. Numerical evaluations by polygraph examiners were unavailable for the validation sample. However, electrodermal response amplitude, cardiograph increase in baseline, and respiration decrease in line length are used by the CPS program, and its unbiased accuracy on the 80 cases in the standardization sample was compared to that of the 32 federal examiners. Excluding 11% inconclusive outcomes, CPS decisions were 90.9% and 83.3% correct for truthful and deceptive cases, respectively. The area under the receiver operating characteristic (ROC) curve was greater for CPS (.919) than for any of the 32 polygraph interpreters (M = .882).

The results from the present study support the following recommendations:

- The Axciton polygraph should be modified or replaced by an instrument that uses scientifically acceptable techniques for recording electrodermal activity. At least one highly diagnostic electrodermal criterion was rendered useless by the Axciton polygraph.

- Decrease in the cardiograph baseline and decrease in respiration baseline should not be considered indications of psychophysiological arousal. The observed relationships with deceptive status were opposite to those predicted by 1999 DoDPI criteria. The present findings support the decision by DoDPI in 2003 to drop these criteria from the curriculum.

- The inter-question interval following a strong cardiovascular response should be increased to a minimum of 35 seconds, or a neutral question should follow a question that evokes a strong cardiovascular response.

- Automated polygraph systems should have the capability to display both unfiltered and high-pass filtered cardiograph to assist in the numerical evaluation of this channel.

- Training should focus on scoring respiration. The reliability and validity of numerical evaluations of respiration would improve if the evaluations were based on actual measurements of line length.

- Most of the DoDPI criteria for respiration may be replaced by line length.

- Arousal is indicated in respiration when a decrease in respiration amplitude or an increase in respiration baseline persists for three or four cycles following question onset. Two-cycle changes are not diagnostic.

Background

Methods for interpreting physiological recordings from comparison question polygraph tests may be broadly classified as global, numerical, or computer methods. Polygraph examiners who use global methods form an overall impression of the strength and consistency of the examinee's physiological reactions to comparison and relevant questions. To decide if the subject was deceptive, the polygraph examiner combines information obtained from the polygraph charts with informal evaluations of the case facts and the subject's statements and demeanor during the polygraph test (Reid & Inbau, 1977).

Polygraph examiners who use numerical methods also assess the relative strength of physiological reactions to comparison and relevant questions, but they do so in a systematic manner according to an established set of rules (e.g., Bell et al., 1999; Swinford, 1999). The rules of numerical scoring specify scoring windows, exclusionary criteria, and types of physiological changes that qualify as reactions. In addition, the

physiological recordings constitute the only source of information that is formally used to reach a decision.

Raskin (1976) compared the decisions made by professional polygraph examiners who used either global or numerical methods for scoring polygraph charts. He provided each of 25 polygraph examiners with the charts from 16 field polygraph examinations. The 16 suspects had been either cleared of wrongdoing by the confession of another person (4 truthful suspects) or incriminated by their own confession (12 deceptive suspects). Eighteen of the polygraph examiners used global methods to evaluate the polygraph charts, and the remaining seven examiners used numerical methods.

The mean accuracy of decisions made by the interpreters who used global methods (M = 87.4%) was significantly lower than the accuracy achieved by the seven numerical evaluators (M = 98.9%). The difference between global and numerical methods was greatest for innocent suspects. The percentage of false positive errors for the global evaluations (26.4%) was over seven times greater than the percentage of false positive errors for the numerical evaluations (3.6%). The results of this study suggest that systematic, rule-bound approaches to chart interpretation are superior to less formal, global methods.

Numerical methods for scoring charts were originally introduced by Backster (1963; 1969). Since then, alternative methods of numerical evaluation were developed by the United States Army (Weaver, 1980; 1985) and the University of Utah (e.g., Podlesny & Raskin, 1978). Although Backster's numerical scoring system represented a major improvement over global approaches to chart interpretation, his scoring rules were numerous, complex, and were not validated by scientific research. A scoring system that consists of many complex rules is not likely to produce reliable outcomes. Numerical scorers who attempt to apply numerous complex rules will evaluate the same set of polygraph charts differently because they focus on different characteristics of the physiological recordings, weigh certain criteria more than others, misapply some of the rules, or simply forget them. Conversely, a scoring system that consists of a few simple rules is likely to result in high levels of agreement among different evaluators (reliability). Although reliability does not guarantee that the judgments will be correct, it does place an upper limit on the validity of decisions. For a scoring system to produce accurate decisions, it must be reliable.

The numerical scoring system introduced by the U. S. Army (Weaver, 1980; 1985) improved upon the Backster method by reducing both the number and complexity of scoring criteria. In 1999, the numerical methods taught at DoDPI consisted of 23 criteria (Swinford, 1999). In 2003, the number of criteria was reduced to 20 (Stern, 2003). Rules for numerical scoring developed independently at the University of Utah further reduced the number of criteria to 10 (Bell et al., 1999). Consistent with the idea that fewer is better, Senter, Dollins, and Krapohl (1999) found that evaluations based on Utah criteria were more reliable and more valid than those based on DoDPI criteria.

Computer methods developed at the University of Utah made it possible to quantify and assess the reliability and diagnostic validity of each of the 10 Utah scoring criteria individually (Kircher & Raskin, 1988). In addition, the reliability and accuracy of diagnoses based on applications of the Utah scoring system has been assessed in a number of laboratory and field studies (Bell et al., 1999).

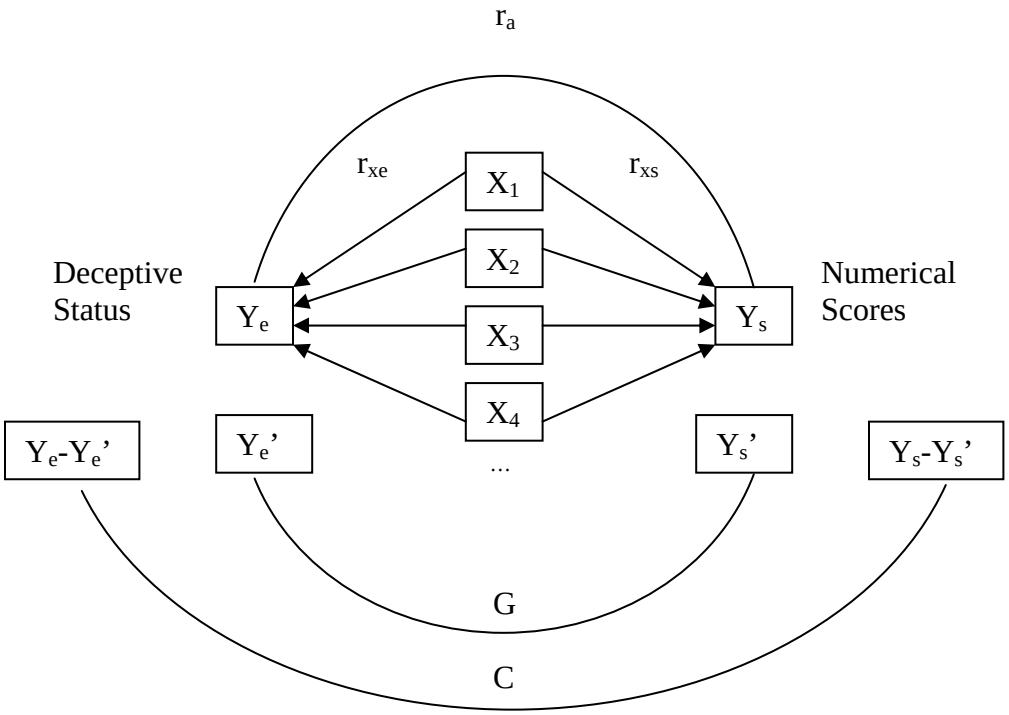
Less is known about the psychometric properties of the criteria used by DoDPI. Harris, Horner, and McQuarrie (2000) explored the ability of each of the DoDPI criteria to distinguish between truthful and deceptive subjects. They concluded that only three of the criteria were sufficient to score charts. The criteria they identified were electrodermal response amplitude, cardiograph change in baseline, and respiration line length. However, the manner by which they measured the diagnostic validity of the various criteria made it difficult to relate their findings to prior research. The present study used traditional psychometric techniques to assess the stability of DoDPI criteria across probable-lie/relevant question pairs (reliability) and to reevaluate their diagnostic validity.

Another objective of the present project was to determine which of the DoDPI criteria are used by federal examiners when they numerically evaluate polygraph charts. We used Brunswik's (1952; 1956) lens model to address this question. The lens model described the judgments by the examiner (numerical scores) and the optimal classification strategy in terms of linear combinations of physiological measures. The lens model organized three sources of information and the relationships among them as illustrated in Figure 1.

As shown on the left side of Figure 1, the statistically optimal classification strategy was defined in terms of a multiple regression equation in which the actual deceptive status of subjects (Y_e) was predicted (Y_e') from a linear combination of physiological measures or cues (X_i). The subscript *e* in the lens model stood for the environment or, in this case, the deceptive status of suspects. The deceptive status of suspects (Y_e) was a dichotomous variable that distinguished between the truthful and deceptive cases in the sample. The simple correlations between the physiological measures and deceptive status (r_{xe}) were cue validity coefficients. These correlations indicated the extent to which the individual physiological measures were useful for discriminating between truthful and deceptive responses to relevant questions. Additionally, the multiple correlation between the set of physiological measures and the criterion (R_e) provided a measure of the validity of the combination of physiological measures for predicting deceptive status.

On the right side of the figure, the decision policy of the polygraph interpreter was represented as the regression of numerical scores (Y_s) for a given physiological channel on the features extracted from that channel. The subscript *s* referred to the polygraph interpreter who served as the

Figure 1. Brunswik's Lens Model



subject of the lens model analysis. The simple correlations between physiological cues and numerical scores (r_{xs}) were cue utilization coefficients. These correlations indicated the extent to which the interpreter's numerical scores depended on each of the physiological measures. These correlations and partial regression coefficients indicated which of the DoDPI scoring criteria the interpreters used when they evaluated the charts. The multiple correlation R_s indicated how well the combination of physiological measures accounted for variance in examiners' numerical evaluations overall.

The correlation between numerical scores (Y_s) and deceptive status (Y_e) was the measure of achievement (r_a) and is the most important component of the lens model. The magnitude of r_a indicated how well the interpreter discriminated between the truthful and deceptive suspects. The matching coefficient, *G*, was the correlation between predicted deceptive status (Y_e') and predicted numerical scores (Y_s'). Conceptually, *G* indicated how closely the interpreter's use of the available physiological measures (his/her decision policy) matched the optimal linear combination of physiological measures.

Finally, the *C* component represented the degree to which nonlinear variance in one system ($Y_e - Y_e$') correlated with nonlinear variance in the other system ($Y_s - Y_s$'). *C* usually is regarded as measure of nonlinear information processing. However, in the present context, the magnitude of *C* also depended on the presence of diagnostic information on the printed chart that was available to the numerical evaluator but was not represented in the discrete computer measurements of the DoDPI criteria. We expected *C* to tell us if the accuracy of decisions by the computer could be improved by modeling the procedures used by polygraph examiners when they score charts.

Szucko and Kleinmuntz (1981) and Kleinmuntz and Szucko (1982) conducted lens model analyses of the decisions made by Reid-trained polygraph examiners and students. However, numerous problems with those studies make it difficult to interpret the meaning of their generally null results (Ben-Shakhar & Lieblich, 1984; Kircher & Raskin, 1983; Kircher, Horowitz, & Raskin, 1988). Subsequently, we used the lens model to compare Utah and Backster numerical scoring techniques, to model the decision policies of professional field polygraph examiners, and to compare the decision policies of polygraph examiners to an optimal linear combination of computer-generated physiological measures (Kircher & Raskin, 1983; Kircher et al., 1995).

One interesting finding that emerged from this research was that some expert polygraph interpreters reliably extracted more diagnostic information from the physiological recordings than is represented in the computer model (Raskin et al., 1988). The *C* coefficient from the lens models for the four federal examiners ranged from .32 to .51, and all were statistically significant. These findings raise the possibility that automated chart analysis programs may be improved by modeling the procedures used by expert polygraph interpreters when they numerically evaluate charts.

The last objective of the research was to develop an optimal combination of DoDPI features that maximizes the separation between confirmed truthful and confirmed deceptive cases. The combination of measures was optimal in the sense that variables were selected and mathematically weighted to minimize decision errors. Two approaches have been used to develop linear combinations of physiological indicators of deception. Kircher and Raskin (1988) used linear discriminant function analysis, first recommended by Kubis (1973). Olsen, Harris, Capps, and Ansley (1997) used logistic regression. The two approaches make different assumptions about the distributions of scores on the variables in the model. However, given the same physiological measures, diagnoses by discriminant functions or logistic regression models are virtually identical (Kircher et al., 2001; Kircher, Raskin, Honts, & Horowitz, 1994).

Methods

Polygraph Cases

The polygraph cases consisted of 80 confirmed truthful and 80 confirmed deceptive field examinations from the database of criminal cases maintained by DoDPI. All of the examinees had been given single-issue Zone Comparison or Modified General

Question Tests (MGQT). Descriptive characteristics of the cases are presented in Table 1. Truthful suspects' veracity had been established by the subsequent confession of another individual that cleared the suspect of any wrongdoing. Deceptive suspects' veracity had been established by their own admission that they had lied to the relevant questions sometime after they had taken the polygraph test or by irrefutable physical evidence (Dollins et al., 1999).

The cases were divided into standardization and validation samples. The standardization sample consisted of 40 truthful and 40 deceptive cases selected at random from the 41 truthful and 56 deceptive cases in the Dollins et al. (1999) study. Three charts of physiological data were available for each case. The standardization sample was used to assess the reliability and validity of DoDPI criteria and to conduct the lens model and discriminant function analyses.

The validation sample consisted of 40 truthful and 40 deceptive cases from the database of verified case maintained by DoDPI. The validation sample was used to test the discriminant functions and assess the stability of results for individual scoring criteria across independent samples.

Polygraph Interpreters

Each of 32 federal polygraph examiners independently evaluated all of the polygraph charts in the standardization sample. Two interpreters (6%) were women. All but one of the interpreters were white. Ages ranged from 27 to 58 years (M = 45.7, SD = 7.5). Education ranged from 14 to 21 years (M = 17.4, SD = 1.6). Fifteen interpreters were affiliated with DoDPI (47%). The remaining 17 interpreters worked for the US Air Force (6), NSA (4), US Army (2), CIA (1), CIFA (1), FBI (1), NRO (1), or US Secret Service (1). Twelve interpreters worked in law enforcement, 19 worked for intelligence agencies, and one was unclassified. Years of experience interpreting polygraph charts ranged from 2 to 23 (M = 6.3, SD = 6.3). Overall, the sample consisted of older, well-educated, and experienced federal polygraph examiners. Other descriptive information is included in Appendix A.

Table 1. Descriptive Characteristics of Sample Cases

	Standardization Sample (N=80)			Validation Sample (N=80)		
	Deceptive	Truthful	Total	Deceptive	Truthful	Total
Test Format						
ZCT	20	22	42	20	13	33
MGQT	19	17	36	20	22	42
U-phase	1	1	2	0	5	5
Total	40	40	80	40	40	80
Field Examiner's Decision						
Deceptive	33	6	39	34	1	35
Nondeceptive	0	25	25	0	32	32
No decision	2	5	7	4	3	7
Unknown	5	4	9	2	4	6
Total	40	40	80	40	40	80
Confirmation Type						
Confession by other	N/A	32	32	N/A	35	35
Examinee confession only	33	0	33	27	0	27
Examinee confession with evidence	2	3	5	3	0	3
Evidence only	1	1	2	10	5	15
Unknown	4	4	8	0	0	0
Total	40	40	80	40	40	80

Type of Crime						
Theft / robbery / larceny	21	24	45	8	30	38
Sexual	7	5	12	12	2	14
Drug offenses	0	0	0	11	0	11
Forgery / false statement	4	0	4	5	3	8
Violent crimes	2	3	5	2	4	6
Physical (nonsexual) child abuse	0	0	0	0	1	1
Unknown	6	8	14	2	0	2
Total	40	40	80	40	40	80
Examinee Type						
Suspect	36	34	70	37	33	70
Witness	3	3	6	0	3	3
Victim	1	3	4	1	1	2
Unknown	0	0	0	2	3	5
Total	40	40	80	40	40	80
Examinee Sex						
Male	24	31	55	35	28	63
Female	15	9	24	3	8	11
Unknown	1	0	1	2	4	6
Total	40	40	80	40	40	80
Examinee Race						
White	22	10	32	9	18	27
Black	14	5	19	3	9	12
Hispanic	2	1	3	3	0	3
Asian	0	0	0	0	1	1
Pacific Islander	1	0	1	0	0	0
Unknown	1	24	25	25	12	37
Total	40	40	80	40	40	80
Examinee Age (years)						
Mean	32.0	31.9		26.4	30.2	
SD	11.4	9.7		7.4	8.9	
Min.	14	17		17	20	
Max.	60	54		49	57	
n	39	37		36	36	
Examinee Education (years)						
Mean	11.5	13.0		12.6	12.7	
SD	2.3	2.0		1.5	2.6	
Min.	7	9		10	3	
Max.	20	16		18	18	
n	36	35		32	34	

Procedures

All of the physiological recordings had been made with Axciton computerized polygraphs (Axciton Systems, Inc., Houston, TX). The Axciton computer files provided continuous recordings of thoracic respiration, abdominal respiration, electrodermal activity, and cardiograph with a minimum inter-question interval of at least 25 seconds. Electronic versions of the Axciton charts were converted from their native binary format to 60 Hz ASCII streams by a computer program entitled Reformat (Version 1.1; Harris & Horner, 1999). The ASCII data were imported into CPS (Version 3.21; Stoelting Company, Wood Dale, IL). Relevant questions were relabeled R1, R2, etc. and probable-lie

comparison questions were relabeled C1, C2, etc. according to their orders of appearance on the chart. The charts then were printed in color. The three charts for each subject were stacked on top of each other, laminated, randomized, and labeled 1 to 80.

Interpreters independently scored the 80 sets of charts in counterbalanced order over a period of 3-5 days at DoDPI in Fort Jackson, SC or at the Defense Security Service Academy facility, Linthicum, MD. Participants were provided informed consent and were assured that their performance would be kept confidential. Participants completed a demographic questionnaire and received general instructions about completing the scoresheets before they began their numerical evaluations. Interpreters were unaware of the deceptive status of tested individuals, and they were unaware of the base rate of deception.

Scoresheets

Numerical scoresheets were created for this project that could be scanned and entered automatically into a computer database (Appendix B). Each interpreter completed one scoresheet for each chart. The interpreter assigned a numerical score that ranged from -3 to +3 for each of the four recorded channels of physiological activity and for each presentation of a relevant question. They also assigned a single numerical score to each relevant question based on their evaluation of the thoracic and abdominal respiration responses combined. The option to not score a response (NS) was provided on the scoresheet, as was space to make comments. Finally, for each numerical evaluation, the interpreter indicated the probable-lie question he or she used as the basis of comparison. As interpreters completed their evaluations of the charts, their scoresheets were checked for completeness and scanned into a database.

Feedback

After the interpreter had completed the 80 sets of charts, an experimenter provided the interpreter confidential and individualized feedback about their performance. Each interpreter was given a binder of computer-generated bar graphs that showed (1) the interpreter's percent correct, wrong, and inconclusive outcomes for deceptive and truthful cases separately; (2) the percent correct and wrong for each physiological channel separately and for the two respiration channels combined; (3) and the interpreter's cue utilization coefficients for each of the DoDPI scoring criteria. Percent correct, wrong, and inconclusive were based on the sum of numerical scores for combined respiration, electrodermal, and cardiograph channels across questions and charts. A total score of +6 or greater was considered a truthful outcome, a total score of -6 or lower was a deceptive outcome, and total scores between the two cutoffs were inconclusive. To obtain dichotomous decisions for each channel, any positive total or zero was considered a truthful outcome, and any negative total was considered a deceptive one.

In each graph, except those containing cue utilizations, the results for the interpreter were shown side-by-side with the results from a computer analysis of the same data. Cue utilizations were shown with cue validities. Cue utilizations were bivariate correlations between computer measurements of each of the DoDPI criteria and the interpreter's numerical evaluations. Cue validities were correlations between computer measurements of the DoDPI criteria and the actual deceptive status of the 80 tested individuals. To facilitate comparisons with the numerical evaluations, the sign of respiration line length was reversed. This strategy ensured that if the predicted difference between responses to probable-lie and relevant questions occurred, the correlation with deceptive status always would be positive and, presumably, in the same direction as the numerical scores (Kircher & Raskin, 2001). An example of the feedback for one interpreter is provided in Appendix C.

Computer Measurements of DoDPI Criteria

ASCII versions of the Axciton charts stored at 60 Hz were imported into CPSLAB Version 7, which was developed for this project to measure the DoDPI criteria described by Swinford (1999).

Electrodermal Measures

The 60 Hz samples of electrodermal data for a chart were transformed to z scores. To be considered an electrodermal response, the amplitude had to exceed a threshold value

of 0.1 standard score units. Changes less than this threshold were considered noise and were not counted as responses. The threshold was established by trial-and-error to approximate distributions of counts of electrodermal responses achieved with a cutoff of 0.02 μS of skin conductance in a previous mock crime experiment (Kircher et al., 2001). Three electrodermal response criteria were extracted from the electrodermal signal that began at question onset and ended 20 seconds later. The electrodermal measures were:

Amplitude. Low points in the electrodermal signal were identified as changes from negative or zero slope to positive slope, and high points in the response curve were identified as changes from positive slope to zero or negative slope. The difference was measured between each low point that occurred prior to the indication that the subject answered and every succeeding high point. Peak amplitude was defined as the greatest observed difference.

Complexity. Complexity was the number of responses that exceeded the 0.1 standard score threshold (peaks) in the 20-second scoring interval.

Duration. Duration was the time from response onset to when the response returned to the level at response onset. Because the Axciton filtered low frequency components from the electrodermal signal, it was rare that the response did not recover fully to the level at response onset. In that event, duration was the time from response onset to the end of the 20-second scoring interval.

Cardiograph Measures

Swinford (1999) described eight cardiovascular criteria, one of which was premature ventricular contractions (PVC). The PVC criterion was not measured in the present investigation due to its low incidence in field data (Ansley & Krapohl, 2000).

To measure the remaining seven criteria, CPSLAB identified the times and levels of systolic and diastolic points in the 60 Hz cardiograph signal that began 5 seconds prior to question onset and ended 20 seconds later. It then calculated a weighted average of the systolic points that occurred during each poststimulus second (Kircher & Raskin, 2001). The resulting series of 25 poststimulus systolic averages defined a systolic response curve. The same procedure was used to create a second-by-second diastolic response curve. A mean response curve was computed by averaging the systolic and diastolic levels for each second. Second-by-second changes in pulse amplitude were obtained by subtracting the diastolic level for each second from the corresponding systolic level. Pulse rate was obtained from the times between systolic points measured to the nearest 17 ms. Second-by-second inter-beat interval curves were obtained by computing a weighted average of the times that occurred during each poststimulus second. From these response curves, CPSLAB extracted the following cardiovascular measures:

Phasic increase in baseline. Points of inflection were identified in the mean cardiograph response curve following question onset. For each low point that occurred prior to the subject's answer, the difference was obtained from every subsequent high point. Phasic increase in baseline was the greatest observed difference.

Tonic baseline increase. Mean cardiograph response levels from 5 seconds prior to question onset to 20 seconds after question onset were correlated with a vector of five zeros followed by 20 ones. If the obtained correlation was negative, it was set to zero. The resulting value was transformed to Z_r using Fisher's r-to-z formula (McNemar, 1968). A large value of Z_r indicated good fit between the data and a model that predicted an abrupt, permanent increase in the cardiograph signal that began at question onset.

Tonic baseline decrease. Mean second-by-second cardiograph levels from 5 seconds prior to question onset to 20 seconds after question onset were correlated with a vector of five zeros followed by 20 negative ones. If the obtained correlation was negative, then it was set to zero. The resulting value was transformed to Z_r using Fisher's r-to-z formula. A large value of Z_r indicated good fit between the data and a model that predicted an abrupt, permanent decrease in the cardiograph signal that began at question onset.

Pulse amplitude increase. The mean of five prestimulus pulse amplitudes was subtracted from each poststimulus pulse amplitude. Pulse amplitude increase was the sum of differences greater than zero.

Pulse amplitude decrease. This measure was obtained in the same manner as pulse amplitude increase except that the pulse amplitudes were reflected (multiplied by -1) prior to accumulating positive differences from the mean prestimulus amplitude.

Inter-beat interval increase (heart rate decrease). This measure was obtained in the same manner as pulse amplitude increase except that second-by-second inter-beat intervals were used instead of mean second-by-second pulse amplitudes.

Inter-beat interval decrease (heart rate increase). This measure was obtained in the same manner as pulse amplitude decrease except that second-by-second inter-beat intervals were used instead of mean second-by-second cardiograph levels.

Respiration Measures

To measure most of the respiration criteria, it was necessary to identify the points of maximum inspiration and maximum expiration for several pre- and post-question onset cycles. For each channel of respiration, the entire respiration signal was transformed to a set of standard scores, smoothed, and sampled at 667 ms to identify approximate times of transitions from negative slope (expirations) to positive slope (inspirations). Approximate times of transitions defined a search space for isolating the precise time and level of a local maximum (peak inspiration) in the original 60 Hz samples. Once identified, local minima (troughs) were isolated between successive times of peak inspiration. Differences between each trough and the subsequent peak were rank ordered. Differences less than 10% of the chart median amplitude were considered too small to be bona fide respiration cycles and were dropped. Cycles also were dropped if differences between successive times of peak inspiration were less than 500 ms in duration. From the times and levels of peaks and troughs in the respiration signal, CPSLAB made the following measurements:

Respiration amplitude was the difference between a trough and the subsequent peak.

Respiration cycle time was the difference between the times of two successive peak inspirations.

I/E ratio was the time from a trough to the next peak (in ms) divided by the time from that peak to the next tough (in ms).

Respiration baseline was the level of a trough in the respiration signal.

Except where noted, two amplitudes, cycle times, I/E ratios, or baseline values that preceded question onset and four values that followed question onset were correlated with a vector of six numbers that represented an idealized pattern of change associated with one of the DoDPI criteria. These models of change are presented in Table 2. Some models were designed to measure changes that lasted for two cycles, whereas others measured changes over three cycles.

The correlation between a model and the respiration data for a test question indicated the extent to which the data fit the pattern of changes specified by a DoDPI criterion. If a pattern of change opposite to the one predicted occurred, it produced a negative correlation that was set to zero to indicate that the predicted pattern was not present. Each correlation then was transformed to Z_r.

Change in respiration was measured relative to a standard. The standard for measuring change in respiration activity was the mean of cycles associated with values of 0 in the measurement models in Table 2. For each model, the first one or two 0's were associated with prestimulus cycles. This approach assumes that the one or two respiration cycles that precede question onset represent 'normal' respiratory activity. A more reliable measure of normal respiratory activity may be the mean amplitude, cycle time, I/E ratio or baseline across the entire chart. To investigate this possibility, each measure described above was recomputed after substituting the chart mean for the measured value for each cycle associated with a model coefficient of 0. The models where the mean

Table 2. Models of Change in Respiration Activity

Model	Description
0, 0, 1, 1, 0, 0	2-cycle increase in amplitude, cycle time, or baseline
0, 0, 1, 1, 1, 0	3-cycle increase in amplitude, cycle time, or baseline
0, 0,-1,-1, 0, 0	2-cycle decrease in amplitude, cycle time, I/E ratio, or baseline
0, 0,-1,-1,-1, 0	3-cycle decrease in amplitude, cycle time, I/E ratio, or baseline
0, 0, 1, 2, 1, 0	Progressive increase in amplitude over 2 cycles followed by a progressive return to baseline
0, 1, 2, 3, 2, 1	Progressive increase in amplitude over 3 cycles followed by a progressive return to baseline[a]
0, 0,-1,-2,-1, 0	Progressive decrease in amplitude over 2 cycles followed by a progressive return to baseline
0,-1,-2,-3,-2,-1	Progressive decrease in amplitude over 3 cycles followed by a progressive return to baseline[a]
0, 0, 1, 2, 0, 0	Progressive increase in amplitude over 2 cycles followed by an abrupt return to baseline
0, 0, 1, 2, 3, 0	Progressive increase in amplitude over 3 cycles followed by an abrupt return to baseline
0, 0,-1,-2, 0, 0	Progressive decrease in amplitude over 2 cycles followed by an abrupt return to baseline
0, 0,-1,-2,-3, 0	Progressive decrease in amplitude over 3 cycles followed by an abrupt return to baseline
0, 0, 1, 1, 1, 1	Permanent increase in respiration baseline
0, 0,-1,-1,-1,-1	Permanent decrease in respiration baseline

[a]One, rather than two, values prior to question onset were used for this model

was substituted for the observed data did not perform as well as those that used the observed prestimulus data and are not discussed further.

Holding and Blocking. To measure *holding* after inhalation or *blocking* after exhalation (Swinford, 1999), the 60 Hz respiration signal was sampled at 4 Hz from question onset for 10 post-question onset seconds. Differences between successive 4 Hz samples were compared to a threshold value that was 20% of the median difference between 4 Hz samples for the entire chart. If the difference did not exceed the threshold and the level of the signal was greater than the mean level of the respiration signal for the chart, it was counted as a 250 ms instance of holding. If the difference did not exceed the threshold and the level of the signal was less than the mean level for the chart, it was counted as an instance of blocking. The holding or blocking measure for a test question was the longest contiguous string of 250 ms instances identified in a 10-second window that began at question onset.

Line length. This was the sum of absolute deviations between successive 60 Hz samples from question onset to 10 seconds after question onset.

Indices of Differential Reactivity to Comparison and Relevant Questions (Cues)

Comparison question techniques predict that truthful suspects will respond more strongly to comparison questions than to relevant questions, whereas deceptive suspects will respond more strongly to relevant questions. An index of differential reactivity to comparison and relevant questions was computed for each subject and

each physiological measure (Kircher & Raskin, 1988). For example, when each of three charts contained two comparison questions and four relevant questions, the computer made 18 measurements of a particular physiological criterion, such as EDR peak amplitude. The 18 measurements were converted to z-scores within the subject. The mean of the 12 z- scores for relevant questions was subtracted from the mean of the 6 z-scores for comparison questions.

The obtained index of differential reactivity is analogous to the total numerical score assigned by the polygraph examiner for a particular channel. The index was positive when the mean reaction to comparison questions was greater than the mean reaction to relevant questions, and the index was negative when the mean reaction to relevant questions was greater. Since truthful suspects were expected to react more strongly to comparison questions and deceptive suspects were expected to react more strongly to relevant questions, we expected positive scores for truthful suspects and negative scores for deceptive suspects. The indices of differential reactivity served as the cues in Brunswik's lens model.

For all variables except respiration line length, a large measured response was indicative of a strong reaction. For line length, suppressed respiratory activity was indicative of a strong reaction. Truthful suspects were expected to show relatively small measured respiration responses (suppression) to comparison questions, whereas deceptive suspects were expected to show relatively small measured respiration responses (suppression) to relevant questions. To achieve a common direction for predicted effects, the sign of the index of differential reactivity for respiration line length was reversed. For every criterion, a positive correlation with deceptive status was expected. A negative correlation would indicate that the observed pattern of response to probable-lie and relevant questions was opposite to the one described in Swinford (1999).

Results

Most of the results from polygraph interpreters are reported in the form of correlation coefficients. As the population correlation approaches a +1 or -1 limit, the distribution of correlations about that population values becomes more skewed. In the present study, we followed McNemar's (1968) recommendation to use Fisher's r-to-z transformation to convert all correlation coefficients to z scores before a mean, variance, or covariance was computed. The resulting statistic in z-score units was then transformed back to the original correlation metric. Unless otherwise noted, all statistical tests were nondirectional and the alpha level was .05.

Reliability of Numerical Evaluations

For each interpreter and each case, the numerical scores assigned to a physiological channel were summed across questions and charts. The sums were arranged in an 80 row (cases) by 32 column (interpreters) matrix. To measure interrater reliability, all possible intercorrelations among the 32 interpreters were obtained. The mean interrater reliability was .89 for the electrodermal channel, .80 for the cardiograph, and .50 for combined respiration. Diagnoses of truth and deception were based on total numerical scores combined across channels. Mean interrater reliability for total numerical scores was .86.

Accuracy of Numerical Evaluations

For each polygraph interpreter, numerical scores for combined respiration, electrodermal, and cardiograph channels were summed across questions, charts, and channels. If the sum was +6 or greater, the subject was classified as truthful. If the sum was -6 or lower, the subject was classified as deceptive. If the total score was between +/-6, the test was considered inconclusive. The percent correct, wrong, and inconclusive for the 40 truthful and 40 deceptive subjects in the standardization sample are presented in Table 3. Also presented in Table 3 is a summary measure of accuracy known as the area under the receiver operating characteristic (ROC) curve (Bamber, 1975). The ROC curve was plotted by varying the numerical cutoff for a truthful/deceptive decision from -108 to +108. Interpreters in Table 3 are ranked in terms of the area under the ROC curve.

Outcomes for an analysis of the same cases by CPS Version 3.21 (Kircher & Raskin, 2001) are presented in the last row. For the analysis by CPS, cases were classified as

Table 3. Percent Decision Outcomes Based on Total Numerical Scores

	Truthful Cases (n=40)			Deceptive Cases (n=40)			
Interpreter	**Correct**	**Wrong**	**Inconc**	**Correct**	**Wrong**	**Inconc**	**ROC**
1	55	8	38	70	5	25	0.918
2	62	17	20	82	5	12	0.912
3	60	12	28	68	0	33	0.911
4	68	8	25	57	12	30	0.904
5	60	5	35	57	3	40	0.901
6	68	10	23	62	3	35	0.901
7	60	5	35	55	5	40	0.900
8	73	5	23	57	8	35	0.899
9	65	8	28	53	5	43	0.897
10	68	5	28	43	10	47	0.897
11	62	8	30	60	3	38	0.897
12	73	8	20	45	10	45	0.895
13	53	10	38	68	3	30	0.895
14	53	8	40	65	5	30	0.894
15	60	5	35	55	5	40	0.892
16	55	10	35	60	5	35	0.890
17	75	8	17	45	8	47	0.886
18	57	10	33	55	0	45	0.886
19	62	8	30	53	3	45	0.885
20	70	12	17	70	5	25	0.884
21	70	12	17	60	5	35	0.878
22	68	10	23	50	5	45	0.878
23	55	8	38	62	3	35	0.875
24	68	8	25	45	12	43	0.873
25	65	8	28	35	5	60	0.872
26	50	20	30	62	3	35	0.854
27	50	8	43	55	5	40	0.853
28	43	15	43	73	5	23	0.849
29	60	10	30	65	12	23	0.841
30	53	15	33	70	10	20	0.838
31	68	5	28	35	15	50	0.838
32	57	15	28	50	10	40	0.837
Mean	**61**	**10**	**30**	**58**	**6**	**37**	**0.882**
CPS	**80**	**8**	**12**	**75**	**15**	**10**	**0.919**

truthful if the probability of truthfulness exceeded .70, deceptive if the probability was less than .30, and inconclusive otherwise. Discriminant scores were used in lieu of numerical scores to calculate the area under the ROC curve.

The inconclusive rate ranged from 17% to 47%. Excluding inconclusives, 85.9% of interpreters' decisions were correct for truthful cases, and 90.6% were correct for deceptive cases. Excluding 11% inconclusive outcomes for CPS, decisions were 90.9% and 83.3% correct for truthful and deceptive cases, respectively. CPS achieved better discrimination between the groups (ROC = .919) than all of the human interpreters

(maximum ROC = .918) and the original examiners (ROC = .908).

Effects of Time on Task

Since interpreters were asked to evaluate a large number of polygraph charts over a few days, fatigue may have had an adverse effect on interpreters' performance. Since the order in which charts were evaluated was counterbalanced over interpreters, we were able to test for effects of time on performance. Growth curve analysis (Raudenbush & Bryk, 2002) was conducted to test for effects of fatigue over days and within days (cf., Lockette & Kircher, 2003).

There was no effect of time-of-day on the accuracy of interpreters' numerical evaluations for deceptive, $t(1227) = 1.59$, $p > .05$, or truthful cases, $t(1227) = -1.55$, $p > .05$. Interestingly, there was improvement in accuracy over days (practice effects). Over a period of four days, the mean numerical score for deceptive cases decreased an average of 1.78 units per day, $t(1227) = -4.04$, $p < .01$, whereas the mean numerical score for truthful cases increased an average of .80 units per day, $t(1227) = 1.88$, $p < .06$.

Objective 1: Assess the reliability and validity of DoDPI criteria for scoring polygraph charts.

Cue Reliability Analyses

To assess the ability of each of the DoDPI criteria (cues) to produce consistent measures of deception across the relevant questions on a test, it was necessary to compute an index of differential reactivity for each relevant question and to use the same number of relevant questions (items) for each subject. Since the numbers of relevant questions on a chart varied from 2 to 4 across cases, only the first two relevant questions were selected for each chart. For the first relevant question on the chart, an index of differential reactivity was computed by subtracting its z-score from the z-score for first probable-lie question on the chart. Similarly, the z-score for the second relevant question was computed by subtracting its z-score from the z-score for the second probable-lie question on the chart. Coefficient alpha then was used to measure consistency across the six presentations of relevant questions on the three charts (Kircher & Raskin, 1988).

The reliability coefficients for the DoDPI criteria are presented in parentheses in Table 4. Statistics are reported separately for thoracic, abdominal, and combined respiration. The combined index of differential reactivity was the mean of the difference scores for thoracic and abdominal respiration. The most reliable measure was electrodermal peak amplitude (r_{xx} = .68). For respiration, reliability was low for all measures except apnea and line length. The reliability of the combined respiration measures generally was slightly greater (M = .21) than the reliability of either thoracic (M = .16) or abdominal respiration (M = .20).

Cue Validity Analyses

Table 4 also contains validity coefficients for each of the DoDPI criteria. For each criterion, the mean of the z-scores for all relevant questions on a test was subtracted from the mean of the z-scores for all comparison questions on the test. The validity coefficient for a criterion was the point-biserial correlation (r_{pb}) between the obtained mean index of differential reactivity and a dichotomous variable that distinguished between truthful (coded 1) and deceptive subjects (coded -1). Correlations above .22 were significantly greater than zero, and correlations below -.22 were significantly less than zero.

As expected, the most valid criterion was EDR peak amplitude (r_{pb} = .67). Of the cardiograph measures, phasic increase in baseline was the most diagnostic criterion (r_{pb} = .51). Of the respiration measures, combined respiration line length was the most diagnostic (r_{pb} = .41).

Importantly, several DoDPI criteria were significantly *negatively* correlated with deceptive status. That is, the presence of these criteria in the physiological data should not be scored, or they should be scored in the opposite manner to that described by Swinford (1999).

Table 4. Validity (and Reliability) of DoDPI Criteria

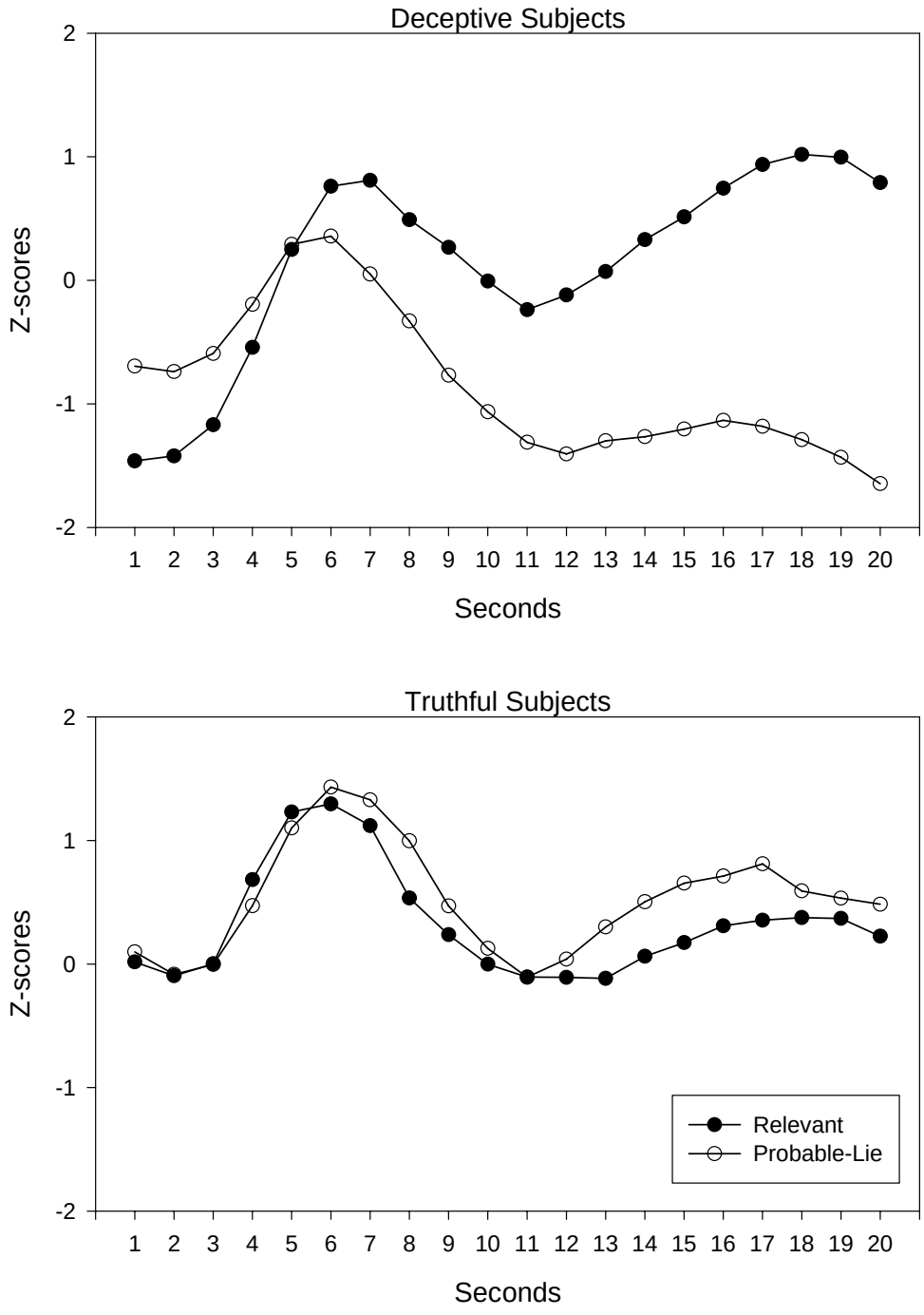
Electrodermal	r_{pb} (r_{xx})
Amplitude	**.67** (.68)
Complexity	-.11 (.14)
Duration	.10 (.26)
Cardiovascular	
Phasic increase in baseline	**.51** (.51)
Tonic increase in baseline	**.48** (.43)
Tonic decrease in baseline	**-.46** (.58)
Pulse amplitude increase	**.31** (.47)
Pulse amplitude decrease	-.16 (.43)
Pulse rate increase	.15 (.22)
Pulse rate decrease	.04 (.27)

Respiration	**Thoracic** r_{pb} (r_{xx})	**Abdominal** r_{pb} (r_{xx})	**Combined** r_{pb} (r_{xx})
Rate abrupt decrease (2 cycles)	-.02 (.06)	**-.23** (.22)	-.14 (.12)
Rate abrupt decrease (3 cycles)	-.03 (.10)	**-.27** (.27)	-.16 (.19)
Rate abrupt increase (2 cycles)	.12 (.01)	.13 (.24)	.14 (.15)
Rate abrupt increase (3 cycles)	.02 (.01)	.04 (.00)	.03 (.00)
I/E ratio abrupt decrease (2 cycles)	-.17 (.03)	-.05 (.13)	.13 (.11)
I/E ratio abrupt decrease (3 cycles)	-.04 (.12)	**-.23** (.04)	-.16 (.11)
Amplitude abrupt increase (2 cycles)	-.05 (.20)	-.06 (.22)	-.06 (.25)
Amplitude abrupt increase (3 cycles)	-.21 (.20)	-.07 (.30)	-.15 (.28)
Amplitude abrupt decrease (2 cycles)	-.14 (.15)	-.12 (.06)	-.15 (.13)
Amplitude abrupt decrease (3 cycles)	.18 (.15)	.21 (.06)	.21 (.17)
Amplitude gradual increase and decrease (2 cycles)	-.22 (.16)	-.22 (.18)	**-.25** (.17)
Amplitude gradual increase and decrease (3 cycles)	-.14 (.25)	-.19 (.17)	-.19 (.23)
Amplitude gradual increase abrupt return (2 cycles)	-.12 (.14)	-.09 (.30)	-.12 (.27)
Amplitude gradual increase abrupt return (3 cycles)	**-.25** (.17)	-.12 (.00)	-.21 (.07)
Amplitude gradual decrease abrupt return (2 cycles)	.08 (.04)	-.11 (.20)	-.10 (.16)
Amplitude gradual decrease abrupt return (3 cycles)	**.33** (.25)	**.27** (.29)	**.33** (.33)
Phasic baseline increase (2 cycles)	.12 (.22)	.02 (.15)	.08 (.19)
Phasic baseline increase (3 cycles)	**.23** (.16)	.18 (.11)	**.23** (.17)
Phasic baseline decrease (2 cycles)	-.09 (.00)	**-.24** (.33)	-.19 (.20)
Phasic baseline decrease (3 cycles)	**-.23** (.00)	**-.27** (.30)	**-.28** (.24)
Tonic baseline increase	**.35** (.25)	**.23** (.29)	**.34** (.38)
Tonic baseline decrease	-.19 (.24)	**-.31** (.36)	**-.30** (.38)
Apnea (holding)	.07 (.27)	.06 (.13)	.09 (.10)
Apnea (blocking)	**.31** (.39)	**.22** (.35)	**.31** (.42)
Line length decrease	**.42** (.50)	**.34** (.27)	**.41** (.43)

Note: Correlations in bold face were significant, $p < .05$.

Invalid cardiograph criteria. For the cardiograph, tonic baseline decrease was negatively correlated with deceptive status (r_{pb} = -.46) and is in conflict with Swinford (1999). To investigate this finding, the cardiograph signal for each chart was converted to z-scores, and second-by-second cardiograph response curves were obtained for deceptive and truthful subjects. The response curves are presented in Figure 2.

Figure 2. Cardiograph Baseline for Deceptive and Truthful Cases



Truthful subjects showed little difference between responses to probable-lie and relevant questions. However, deceptive subjects produced stronger responses to relevant questions. For deceptive subjects, cardiograph responses to relevant questions had not recovered fully to baseline within 20 seconds of question onset. In addition, there was a noticeable drop in baseline following the onset of probable-lie questions. Unfortunately, on 65% of occasions that a probable-lie question was presented to a deceptive subject, it was immediately preceded by a relevant question. Since deceptive subjects' cardiograph responses did not recover prior to the 20th poststimulus second, it appears that the drop in baseline associated with probable-lie questions was part of the recovery from the response to the preceding relevant question, and the negative correlation between deceptive status and baseline decrease is an artifact of question placement.

Invalid respiration criteria. Negative correlations with deceptive status were obtained for decreases in respiration baseline following question onset. Again, the thoracic and abdominal respiration signals for each chart were each converted to z-scores. Baseline values (troughs) for comparison and relevant questions were obtained for 2 pre- and 5 poststimulus cycles. Since the response patterns were similar for thoracic and abdominal respiration, a mean cycle-by-cycle curve was obtained for each subject. Mean differences from the baseline just prior to question onset (cycle 2) are presented in Figure 3.

On average, deceptive subjects showed a reduction in respiration baseline in response to probable-lie questions and little change in response to relevant questions. For truthful subjects, there was little difference in baseline responses to probable-lie and relevant questions. That baseline responses to probable-lie questions were offset by a cycle might be related to differences in lengths of test questions. Probable-lie questions often are longer than relevant questions, and they may have longer response latencies.

Deceptive subjects showed a greater reduction in respiration baseline in response to probable-lie questions than to relevant questions. This finding is opposite to the one predicted by 1999 DoDPI rules and resulted in negative correlations with deceptive status (phasic baseline decrease r_{pb} = -.28; tonic baseline decrease r_{pb} = -.30). The positive correlations for phasic (r_{pb} = .23) and tonic baseline increases (r_{pb} = .34) were consistent with DoDPI rules.

A significant negative correlation also was obtained for thoracic respiration amplitude gradual increase (r_{pb} = -.25). However, the effect was not significant for abdominal respiration or when the thoracic respiration data were combined with the abdominal data.

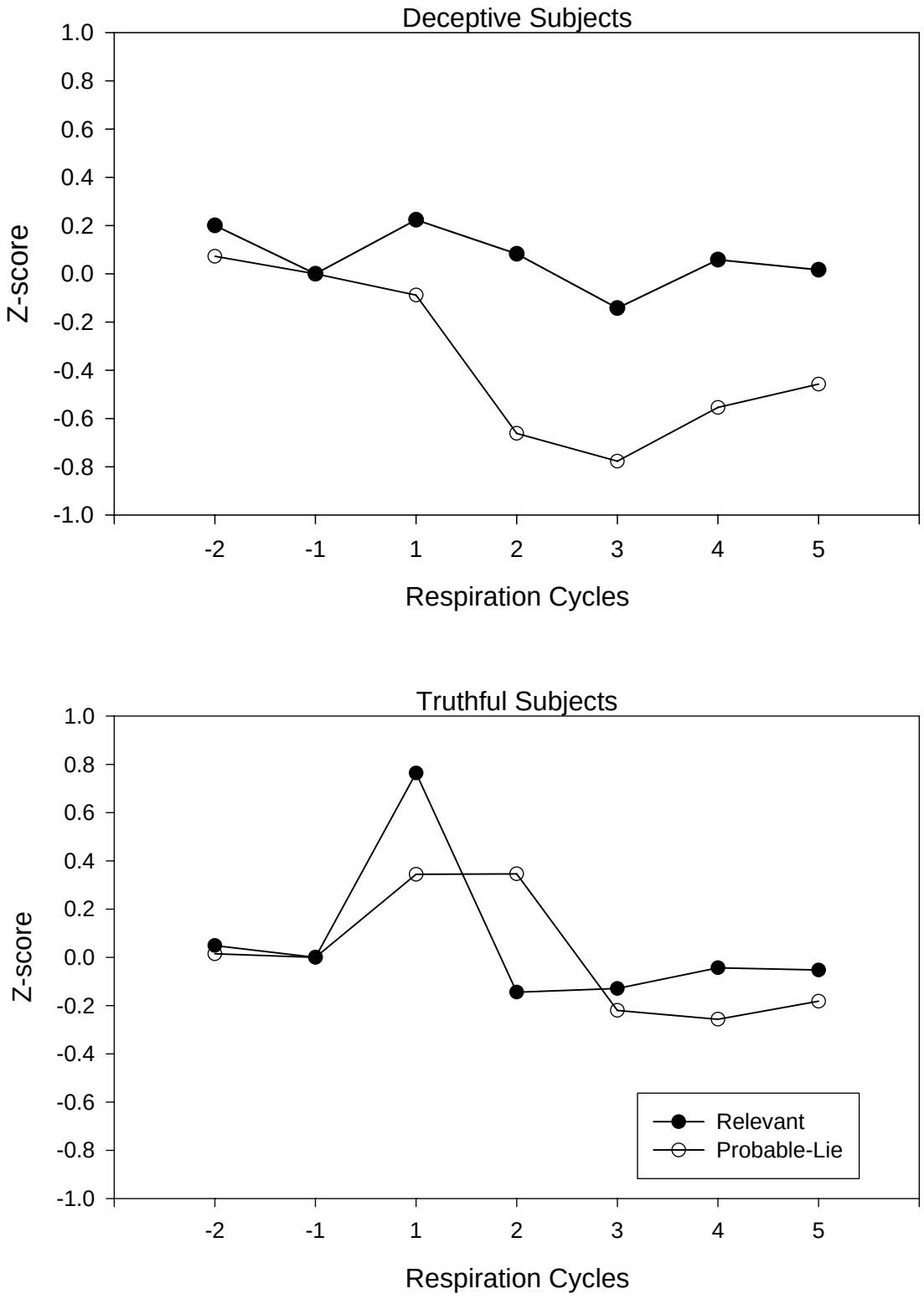
Conditional Applications of Physiological Criteria

DoDPI guidelines recommend that EDR duration and EDR complexity be used only as tiebreakers when there is no visually discernable difference in EDR amplitude between the probable-lie and relevant question. Similarly, changes in pulse amplitude or heart rate are used only as tiebreakers when there is no apparent difference in cardiograph baseline increase. Additional analyses were conducted to investigate the possibility that these criteria are valid when used only as tiebreakers. All of the data for the EDR or cardiograph channel were transformed to standard scores. A just noticeable difference in amplitude was identified in a chart, and the observed difference in standard score units was used as a criterion for selecting all comparison-relevant response pairs with absolute differences less than this criterion. Few of the response pairs in the standardization sample met this criterion. To increase the sample size, standardization and validation samples were combined.

Fifty-four of 960 question pairs showed no discernable difference in EDR amplitude. In this sample of 54 comparisons (28 truthful, 26 deceptive), the point-biserial correlation with deceptive status (validity) was -.01 for EDR duration and .00 for EDR complexity, neither of which was significant.

For the cardiograph, 73 comparisons were selected with no discernable difference in baseline increase between responses to probable-lie and relevant questions (37 truthful and 36 deceptive). In the 73 cases where a tiebreaker was needed, correlations

Figure 3. Cycle-by-cycle Respiration Baseline for Deceptive and Truthful Cases



with deceptive status were .20 for cardiograph pulse amplitude increase, -.16 for pulse amplitude decrease, -.09 for heart rate increase, and .13 for heart rate decrease. None of these correlations was significant.

Summary

EDR amplitude was valid for discriminating between truthful and deceptive cases. EDR complexity and EDR duration were not.

For the cardiograph, phasic increase in baseline, tonic increase in baseline, and pulse amplitude increase were consistent with DoDPI scoring rules. Deceptive subjects showed decreases in cardiograph baseline to probable-lie questions that appeared to be part of the recovery from strong responses to the relevant questions that preceded them. Although this finding contradicts Swinford (1999), decrease in cardiograph baseline was dropped as a DoDPI criterion in 2003 (Stern, 2003). The remaining three criteria were uncorrelated with deceptive status.

For respiration, amplitude gradual decrease, baseline increase, and line length decrease were significantly related to deceptive status and were consistent with DoDPI rules. Phasic and tonic decreases in respiration baseline also were related to deceptive status, but the pattern of changes was contrary to Swinford (1999). In 2003, decrease in respiration baseline was dropped as a criterion (Stern, 2003). Changes in respiration rate and I/E ratio were unreliable and invalid.

Objective 2: Identify physiological criteria used by federal examiners to evaluate polygraph charts.

Cue Utilization

Cue utilizations were bivariate correlations between computer measures of the DoDPI criteria for a physiological channel and the polygraph interpreter's numerical evaluations of that channel. Cue utilization was assessed at the level of the individual relevant question by correlating each individual numerical score with the associated computer index of differential reactivity (cue). For each interpreter, cue utilization measured at the level of individual questions was based on 786 paired observations (80 cases X 3 charts X 3.3 relevant questions per chart). In addition, cue utilization was assessed at the level of the case by calculating the total numerical score for a channel and correlating that total with the associated mean index of differential reactivity. Cue utilization measured at the level of the case was based on 80 paired observations, since there were 80 cases.

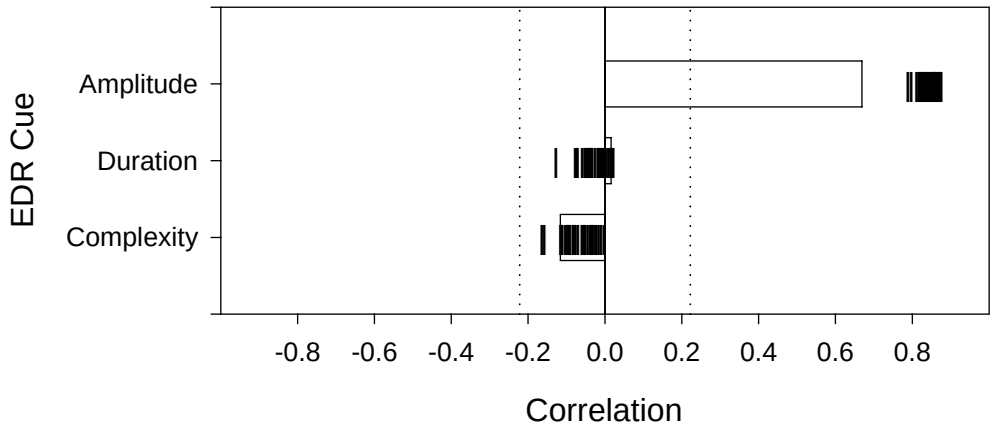
These two approaches yielded similar measures of cue utilization. For each interpreter, cue utilization was measured at the level of the question and again at the level of the case for each of 60 physiological criteria (3 electrodermal + 7 cardiovascular + 25 thoracic respiration + 25 abdominal respiration). The correlation between the 60 paired observations was obtained for each interpreter. The mean correlation between the two measures of utilization was .89. However, cue utilizations were generally greater in absolute magnitude at the level of cases, and only those results are reported.

Utilization of Electrodermal Cues.

Distributions of cue utilizations are presented in Figure 4 for the three electrodermal criteria. Each short vertical line segment shows the cue utilization for an individual polygraph interpreter. The bars in Figure 4 show cue validities (from Table 4) rank ordered from the most positive to the most negative. Cues at the top of the figure are most consistent with the DoDPI scoring rule, whereas cues at the bottom of the figure are the most incongruent with the DoDPI scoring rule. The dotted lines show the cutoffs for statistical significance at $p < .05$. Cues with validity coefficients greater than .22 support the DoDPI scoring rules, whereas cues with validity coefficients less than -.22 are significantly incompatible with DoDPI scoring rules. Null results for a criterion suggest that it may be omitted from the list of DoDPI scoring criteria.

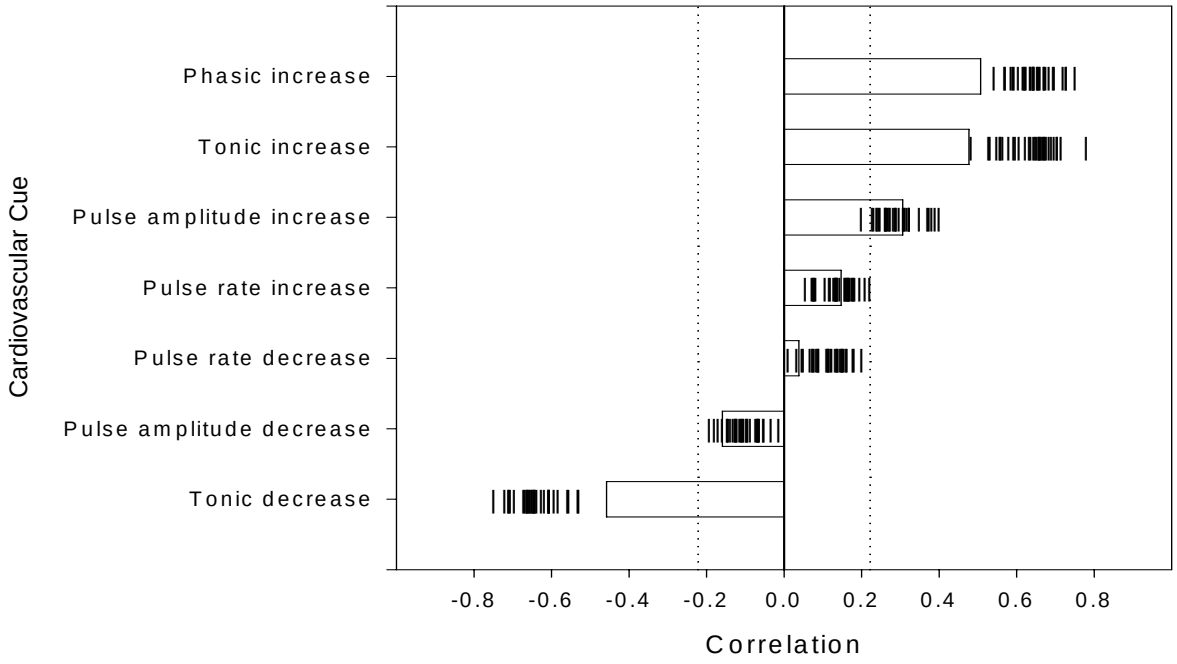
The data in Figure 4 indicate that numerical evaluations of electrodermal responses by every interpreter were significantly and exclusively correlated with EDR amplitude.

Figure 4. Cue validities and Cue Utilizations (|) for Electrodermal Criteria.



Utilization of Cardiovascular Cues. Distributions of cue utilizations are presented in Figure 5 for the seven cardiovascular criteria.

Figure 5. Cue validities and Cue Utilizations (|) for Cardiovascular Criteria.



Interpreters' numerical evaluations of the cardiograph were most highly correlated with phasic (M_r = .65) and tonic (M_r = .64) increases in baseline. Numerical scores also were correlated with pulse amplitude increase (M_r = .30). It appeared that interpreters had negative cue utilizations for tonic decrease. However, negative cue utilizations tonic decrease occurred because interpreters relied on baseline increases and baseline increases were negatively correlated with tonic decrease (r = -.81 for phasic increase; r = -.79 for tonic increase). In multiple regression analyses described below, effects of phasic and tonic increase were controlled, and the relationship between numerical scores and tonic decrease were significant for only 4 (12.5%) interpreters (see Table 8, pg. 102).

Utilization of Respiration Cues. Cue utilizations are presented in Figure 6 for the thoracic, abdominal, and combined respiration criteria. Cues are ranked based on cue validities for the combined respiration criteria. Interpreters relied primarily on line length, amplitude decrease, and apnea. Numerical evaluations by most interpreters were not correlated with baseline permanent increase, even though it was the second most diagnostic respiration measure.

Numerical scores were negatively correlated with increases in respiration amplitude (amplitude abrupt increase, amplitude gradual increase and decrease, and amplitude gradual increase abrupt return). However, all of the measures with the largest positive utilizations (line length, amplitude gradual decrease abrupt return, and amplitude abrupt decrease) were negatively correlated with all of the measures of increases in respiration amplitude. For the combined respiration measures, partial correlations were computed to control for the effects of the three cues with positive utilizations on the relationships between increases in respiration amplitude and numerical scores. Only 5 of 96 partial correlations between increases in amplitude and numerical scores were significant (3 measures of amplitude increase X 32 interpreters). Thus, the observed negative utilizations may be explained by the priority interpreters gave to line length, amplitude gradual decrease abrupt return, and amplitude abrupt decrease. In general, Figures 4, 5, and 6 indicate that there was more variance among interpreters in their utilization of respiration cues than cardiograph or electrodermal cues. These findings are consistent with the interrater reliabilities reported for respiration, cardiograph, and electrodermal activity reported above.

Lens Models

Five lens models were developed for each of the 32 polygraph interpreters. Separate lens models were developed for electrodermal, cardiograph, thoracic respiration, abdominal respiration, and combined respiration channels.

Electrodermal. For the electrodermal channel, deceptive status was regressed onto EDR amplitude, EDR complexity, and EDR duration. The multiple correlation, R_e, was .67. Predicted deceptive status and the residuals from that regression analysis were retained to measure matching (*G*) and response nonlinearity (*C*). Accuracy was assessed for each polygraph interpreter by correlating deceptive status with numerical scores for the EDR channel (r_a). Response linearity (R_s) was assessed by regressing EDR numerical scores onto computer measures of EDR amplitude, EDR complexity, and EDR duration. The predicted values from that regression analysis also were retained. The correspondence between the regression models for deceptive status and for numerical scores (matching) was assessed by correlating the predicted values from the two multiple regressions, *G*. The ability of interpreters to use the physiological cues in a nonlinear manner or to use information in the signal not captured by computer measures of three DoDPI criteria was assessed by correlating the two sets of residuals, *C*. Descriptive statistics for 32 lens model analyses are presented in Table 5.

Figure 6a. Cue validities and Cue Utilizations (|) for Combined Respiration.

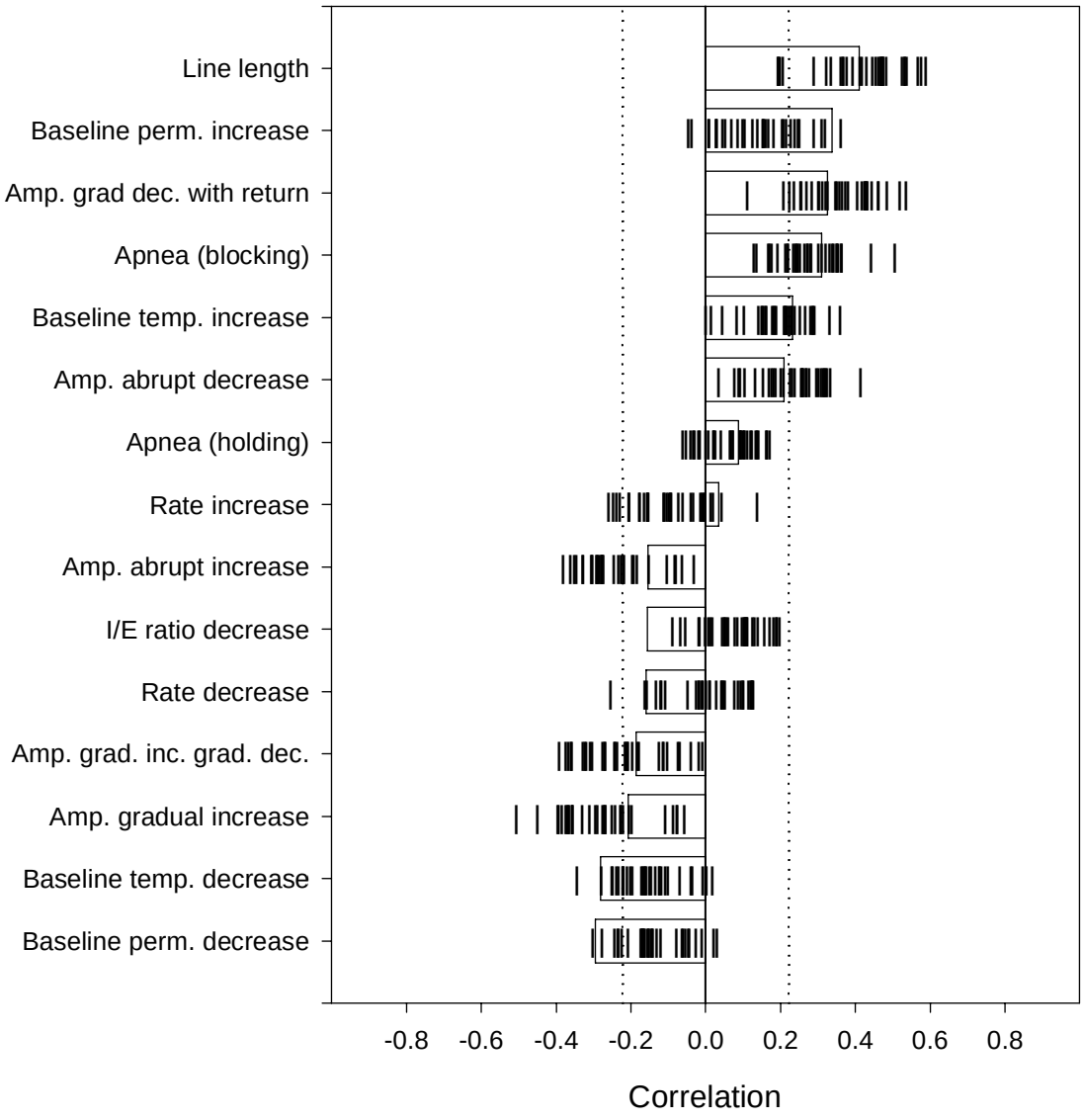


Figure 6b. Cue validities and Cue Utilizations (|) for Thoracic Respiration.

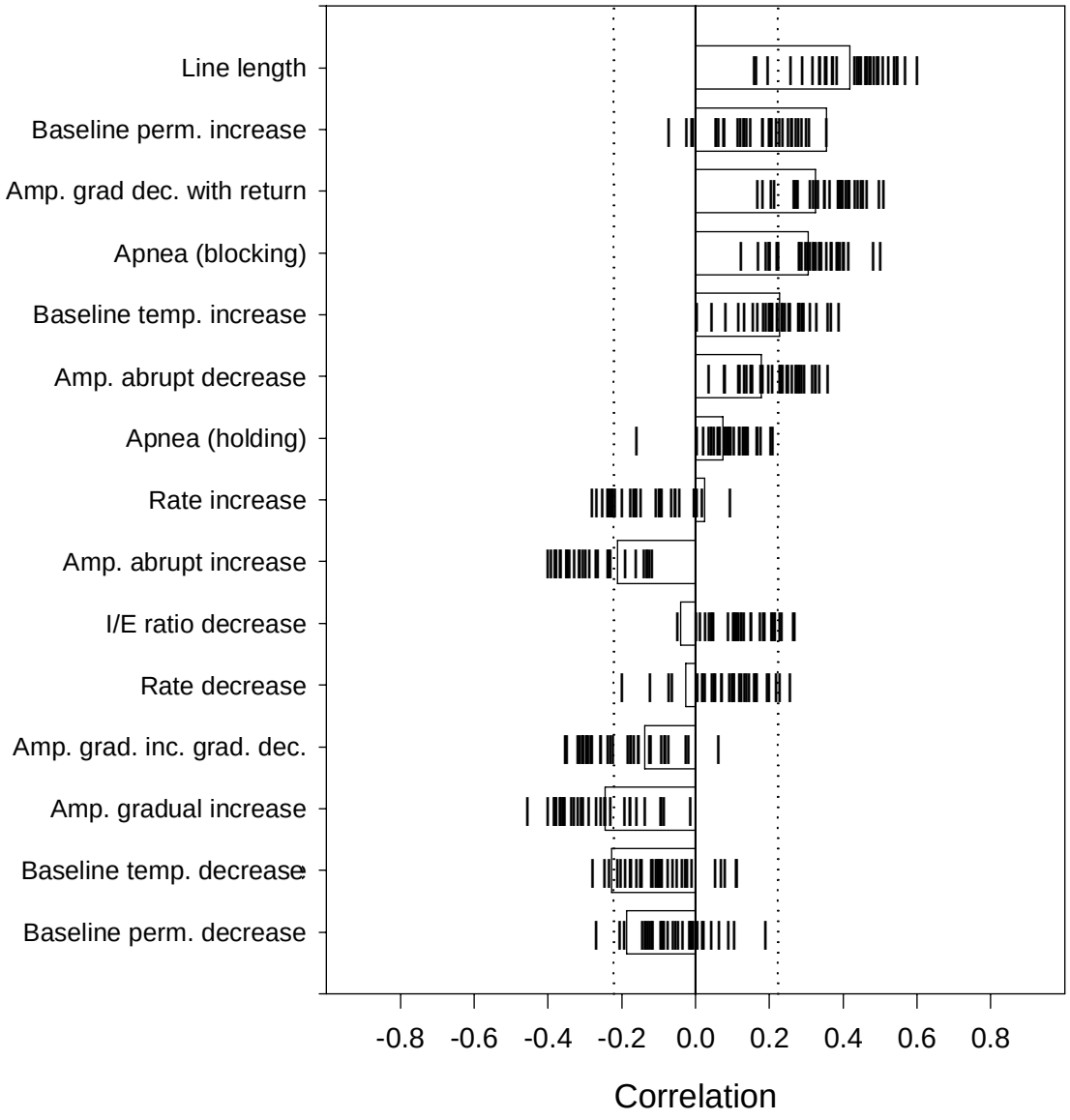


Figure 6c. Cue validities and Cue Utilizations (|) for Abdominal Respiration.

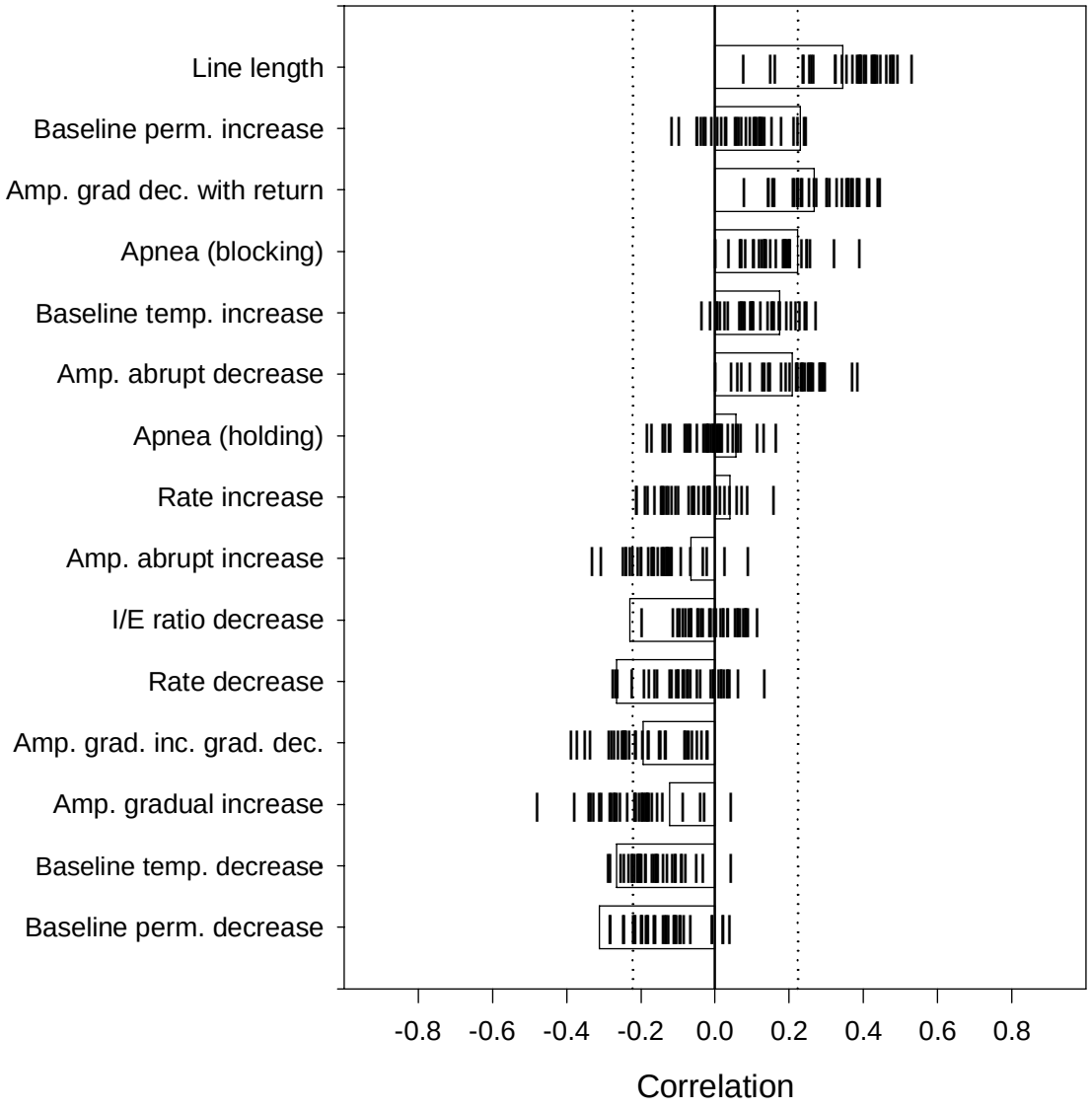


Table 5. Lens Model for Electrodermal Activity (N=32 Interpreters)

Lens Model Component	Mean	SD	Percent
Accuracy (r_a)	.60	.05	100
Response linearity (R_s)	.85	.08	100
Matching (G)	.99	.24	100
Response nonlinearity (C)	.12	.08	16
Cue validity (R_e)	.67		

Note: Percent was the percentage of the 32 interpreters with significant coefficients.

The correlation between numerical evaluations of electrodermal activity and deceptive status (r_a) varied about the mean of .60. 100% of the interpreters provided numerical scores for electrodermal activity that significantly discriminated between truthful and deceptive cases. The optimal linear combination of computer measures of electrodermal activity correlated R_e = .67 with deceptive status. This was one point higher than the maximum value of r_a achieved by any of the numerical evaluators, but it was no better than the simple correlation between EDR amplitude and deceptive status (r_{pb} = .67). Interpreters' numerical scores were predictable from EDR amplitude, complexity, and duration and varied little about the mean R_s = .85. The regression equations for predicting numerical evaluations of electrodermal activity closely matched the optimal linear combination of electrodermal measures for predicting deceptive status, M_G = .99. Finally, there was little evidence that interpreters extracted more useful information from the electrodermal recordings than was included in the computer measures, M_C = .12.

Standardized regression coefficients for predicting EDR numerical scores from EDR cues are summarized in Table 6, along with the optimal weights for predicting deceptive status. As expected, EDR amplitude was the most heavily weighted variable in the regression equations for predicting numerical scores and for predicting deceptive status. All of the interpreters used EDR amplitude, and 6% of the interpreters also used complexity.

Table 6. Standardized Regression Coefficients (B) for Predicting Electrodermal Numerical Scores and Deceptive Status from Electrodermal Cues (N=32 Interpreters)

Electrodermal Cue	Mean	SD	Percent	Deceptive Status
Amplitude	.84	.02	100	.69**
Complexity	-.04	.04	6	.05
Duration	-.03	.03	3	.01

Note: Percent was the percentage of the 32 interpreters with significant coefficients.
*p < .05. ** p < .01.

Cardiograph. Lens model analyses of cardiograph numerical scores, deceptive status, and cardiograph cues are summarized in Table 7.

The optimal linear combination of cardiograph cues (R_e = .57) exceeded the mean accuracy of the interpreters (M_{ra} = .51), although 4 of the 32 interpreters performed as well or better than the optimal combination of cues. Interpreters' numerical evaluations were predictable from cardiograph cues (M_{Rs} = .73), and there was a close correspondence between the regression weights for predicting

deceptive status and the regression weights for predicting numerical scores (M_G = .97). Eight of the 32 interpreters (25%) had statistically significant *C* coefficients (*C* > .22). The latter findings suggest that some polygraph interpreters may extract diagnostic information from the cardiograph recordings that is not adequately represented in the computer measures of the DoDPI criteria.

Table 7. Lens Model for the Cardiograph (N=32 interpreters)

Lens Model Component	Mean	SD	Percent
Accuracy (r_a)	.51	.06	100
Response linearity (R_s)	.73	.12	100
Matching (G)	.97	.25	100
Response nonlinearity (C)	.19	.07	25
Cue validity (R_e)	.57		

Note: Percent was the percentage of the 32 interpreters with significant coefficients.

Standardized regression coefficients for predicting numerical evaluations of the cardiograph are summarized in Table 8, along with the optimal weights for predicting deceptive status. Because the cardiovascular measures were intercorrelated, no variable contributed uniquely to the equation for predicting deceptive status from the seven cardiovascular cues. However, backward elimination yielded a regression equation with two significant predictors, phasic increase in baseline (B = .47, p < .01) and pulse amplitude increase (B = .22, p < .05), and little loss in predictive power, R_e = .55.

Tonic and phasic increase in baseline were the cues most often used by the interpreters to assign their numerical scores. Decreases in baseline, changes in pulse amplitude, and changes in pulse rate did not contribute to the computer model for predicting deceptive status. In only a few cases did these measures contribute to the regression models for predicting numerical scores.

Table 8. Standardized Regression Coefficients for Predicting Cardiograph Numerical Scores and Deceptive Status from Cardiograph Cues (N=32 Interpreters)

Cardiovascular Cue	Mean	SD	Percent	Deceptive Status
Phasic increase in baseline	.27	.11	44	.30
Tonic increase in baseline	.25	.11	53	.20
Tonic decrease in baseline	-.19	.09	13	-.04
Pulse amplitude increase	.15	11	9	.16
Pulse amplitude decrease	.00	.09	0	-.04
Pulse rate increase	.10	.07	3	.06
Pulse rate decrease	.04	.07	0	-.00

Note: Percent was the percentage of the 32 interpreters with significant coefficients.

Respiration. Seventy-five respiration criteria are listed in Table 4. As expected, combined respiration cues were highly correlated with the thoracic cues and with the abdominal cues; thoracic cues were highly correlated with the abdominal cues; and 2-cycle cues were often highly correlated with the corresponding 3-cycle cues. To reduce

multicollinearity and stabilize the regression weights in the lens models, it was necessary to identify a small subset of relatively independent respiration cues. Since the 3-cycle measures tended to be more reliable and valid than the 2-cycle measures (Table 4), we dropped the 2-cycle measures.

To further reduce the number of variables, principal components analyses (PCA) with varimax rotation were performed separately for thoracic, abdominal, and combined respiration cues. For each set of respiration cues, PCA identified four orthogonal components. One component reflected changes in respiration *amplitude*. Cues that loaded on this component included amplitude abrupt increase, amplitude abrupt decrease, amplitude gradual increase and gradual decrease, amplitude gradual increase, and amplitude gradual decrease with return to homeostasis. Another component reflected changes in respiration *baseline*. The measures that loaded on this component included baseline temporary increase, baseline temporary decrease, baseline permanent increase, and baseline permanent decrease. The third component reflected changes in respiration *rate*. It consisted of rate decrease, rate increase, and I/E ratio decrease. The fourth component reflected changes in *apnea*. This component was composed of holding, blocking, and line length. Although line length also loaded on the amplitude and rate components, it loaded more heavily on the apnea factor. For each component, one variable with a salient loading was selected to represent the component and to capture a relatively unique source of variance in the respiration recordings. For each respiration channel, we retained rate decrease, amplitude abrupt decrease, tonic baseline increase, and line length.

A lens model analysis was performed using the four thoracic respiration cues and the numerical scores for the thoracic channel. Another lens model was developed using the four abdominal cues and the numerical scores for the abdominal channel. Finally, a third lens model was developed using the four combined respiration cues and the numerical scores for combined respiration. The results are presented in Table 9.

Table 9. Lens Models for Respiration (N=32 Interpreters)

Lens Model Component	Mean	SD	Percent
Thoracic Respiration			
Accuracy (r_a)	.29	.12	75
Response linearity (R_s)	.47	.12	97
Matching (G)	.84	.36	100
Response nonlinearity (C)	.13	.10	19
Cue validity (R_e)	.50		
Abdominal Respiration			
Accuracy (r_a)	.28	.11	69
Response linearity (R_s)	.45	.11	97
Matching (G)	.84	.32	100
Response nonlinearity (C)	.12	.11	16
Cue validity (R_e)	.51		
Combined Respiration			
Accuracy (r_a)	.30	.12	75
Response linearity (R_s)	.48	.12	97
Matching (G)	.83	.33	100
Response nonlinearity (C)	.12	.10	16
Cue validity (R_e)	.53		

Note: Percent was the percentage of the 32 interpreters with significant coefficients.

The numerical scores assigned to thoracic and abdominal respiration were similar. Within-interpreter correlations between total numerical scores assigned to thoracic and abdominal respiration ranged from .79 to .97 (M_r = .92, SD = .04). Not surprisingly, the lens model results also were similar for thoracic, abdominal, and combined respiration. Since there appeared to be no conceptual or empirical advantage in maintaining separate measures of thoracic and abdominal respiration, only combined respiration measures were retained for subsequent analyses.

Paired-sample t-tests revealed that numerical scores for combined respiration were less valid (M_{ra} = .30) than were numerical scores for the electrodermal (M_{ra} = .60), t(31) = 16.3, p < .01, or cardiograph channels (M_{ra} = .51), t(31) = 12.9, p < .01. In addition, it was more difficult to predict numerical evaluations of respiration (M_{Rs} = .46) from respiration cues than to predict electrodermal scores (M_{Rs} = .85) from electrodermal cues, t(31) = 28.8, p < .01, or to predict cardiograph scores from cardiograph cues (M_{Rs} = .73), t(31) = 21.3, p < .01.

Table 10 contains mean regression weights for predicting combined respiration numerical scores from the four selected DoDPI criteria, as well as the optimal weights for predicting deceptive status. The numerical scores assigned by 88% of the interpreters were significantly correlated with line length. The numerical scores assigned by 5 (16%) of the interpreters were correlated both with line length and tonic baseline increase. These were the two most diagnostic respiration criteria (Table 3). They also were the only two cues that made significant unique contributions to the optimal computer model.

Table 10. Standardized Regression Coefficients for Predicting Combined Respiration Numerical Scores and Deceptive Status from Combined Respiration Cues

Combined Respiration Cue	Mean	SD	Percent	Deceptive Status
Rate abrupt decrease (3 cycles)	-.00	.10	3	-.18
Amplitude abrupt decrease (3 cycles)	.12	.09	9	.03
Tonic baseline increase	.07	.11	16	.26*
Line length decrease	.37	.11	88	.38**

Note: Percent was the percentage of the 32 interpreters with significant coefficients.
*p < .05. ** p < .01.

Individual Differences. To determine if characteristics of interpreters were related to the accuracy of their numerical evaluations, we correlated age, sex, education, affiliation, assignment, and experience with the accuracy (r_a) of numerical scores assigned to electrodermal, cardiovascular, and combined respiration channels. There was little variance in sex since only 2 of 32 interpreters were women. Affiliation was a dichotomous variable that distinguished between interpreters who worked in law enforcement, coded 1 (n=12) or intelligence, coded 0 (n=19). One interpreter was affiliated with neither law enforcement nor intelligence. Assignment was a dichotomous variable that distinguished between interpreters who were or were not assigned to DoDPI. To measure experience, we had planned to pool responses to two questions: "How long have you been conducting polygraph examinations?" and "How long have you been evaluating polygraph charts?" The responses to these items were nearly identical (r = .996). Since the combination of the two items would be no more informative than either item, we chose responses to "How long have you been evaluating polygraph charts?" to represent experience.

Two correlations of interest were statistically significant. Experience correlated .40 with accuracy of numerical scores assigned to electrodermal responses. Interpreters with more years of experience made more accurate numerical evaluations of the electrodermal channel. Experience did not

correlate with cardiograph or respiration scores. In addition, there was a .38 correlation between the accuracy of numerical evaluations of the cardiograph and respiration channels. Interpreters who were better able to extract diagnostic information from the cardiograph also were better able to extract diagnostic information from the respiration signals.

Summary

Results of the lens model analyses suggest that all of the interpreters used EDR amplitude to assign numerical scores to the EDR channel, 2 (6%) also used complexity, and 1 (3%) used EDR duration.

The lens model analyses of the cardiograph suggest that interpreters based their numerical scores on phasic increase in baseline, tonic increase in baseline, and to a lesser extent, tonic decrease in baseline and increase in pulse amplitude. Phasic and tonic increases in baseline and increases in pulse amplitude were consistent with Swinford (1999), but decreases in baseline were not. Decreases in baseline were observed to probable-lie questions when the subject was deceptive to the relevant question. When interpreters used decrease in baseline, they used it appropriately and did not treat it as an indication of psychophysiological arousal.

Numerical scores for thoracic and abdominal respiration were highly correlated and were replaced by numerical evaluations of the two channels in combination. Principal components analysis (PCA) revealed four sources of variance in the respiration measures: changes in rate, amplitude, baseline, and apnea. Line length loaded primarily on the apnea component but also loaded on rate and amplitude components. Decrease in rate, decrease in amplitude, increase in baseline, and decrease in line length were selected to represent the four components. The remaining measures were dropped from further consideration.

Lens analyses of respiration data suggest that numerical evaluations by almost 90% of the interpreters depended on line length. Sixteen percent depended on both line length and tonic baseline increase. Of the four respiration criteria, line length and baseline increase were the only two variables to make unique contributions to the optimal linear model for diagnosing truth and deception. Experience correlated positively with the accuracy of numerical evaluations of electrodermal responses. Finally, interpreters who were better able to score the cardiograph also were better able to score the respiration recordings.

Objective 3: Develop an optimal linear combination of DoDPI features that maximizes the separation between confirmed truthful and deceptive cases.

Eighty-five computer measurements of the DoDPI criteria were available to develop a statistical model for discriminating between truthful and deceptive cases. To improve the chances that the model would generalize to an independent sample of field cases, we identified a subset of variables that met certain criteria. First, each variable had to have a significant bivariate relationship with deceptive status. Second, each variable had to have an internal consistency reliability that exceeded .33. Third, there had to be evidence, independent of the standardization and cross-validation samples, that the variable was diagnostic of truth and deception. To meet the third criterion, we examined the scientific literature. If there were no prior empirical evaluations of the variable, we assessed its diagnostic validity in a sample of 84 subjects from a previous laboratory experiment. The 84 subjects comprised the positive and neutral feedback, probable-lie conditions in Kircher, Packard, Bell, and Bernhardt (2001).

One of the variables that met these requirements was cardiograph tonic decrease in baseline (field r_{pb} = -.46; lab r_{pb} = -.42). Examination of second-by-second plots of the cardiograph responses to probable-lie and relevant questions in the field and laboratory data sets suggested that the drop in cardiograph baseline was due to a strong response to a preceding question that had not fully recovered. That is, the drop in cardiograph baseline was not a response to the question; rather it was a recovery of the response to the preceding question. It seemed unwise to include a variable in the model that depended on the response to the preceding question, especially since the order of question presentation is not standard in the field. Therefore, although it was diagnostic,

cardiograph tonic decrease in baseline was dropped from the final subset of potential predictor variables. In the end, five variables remained. They are listed in Table 11, along with validity coefficients (r_{pb}) from the standardization sample and from the laboratory experiment. The last column in Table 11 shows the r_{pb} for the field validation sample.

.

Table 11. Validity of Cues Selected for All-Possible-Subsets Regression in Field Standardization, Laboratory, and Field Validation Samples

	Standardization N=80 r_{pb}	Laboratory N=84 r_{pb}	Validation N=80 r_{pb}
Electrodermal			
Amplitude	**.67**	**.77**	**.53**
Cardiovascular			
Phasic increase in baseline	**.51**	**.50**	**.45**
Tonic increase in baseline	**.48**	**.45**	.19
Respiration			
Apnea (blocking)	**.31**	**.27**	.12
Line length decrease	**.41**	**.49**	**.30**

Note: correlations in bold were statistically significant at $p < .05$.

When there are only two groups, discriminant analysis and multiple regression are mathematically equivalent. In multiple regression, the dependent measure is a dichotomous variable that represents group membership (deceptive status), and the predictor variables are the physiological cues. Three alternative subsets of the five physiological cues were obtained from an all-possible-subsets regression analysis of the standardization sample (BMDP program 9R, 1993). For each subset, a discriminant function was developed using the standardization cases. The discriminant function was then used to compute the probability of group membership for each case in the validation sample (Kircher & Raskin, 2001). Validation cases were classified as truthful if the probability of truthfulness exceeded .70, deceptive if the probability of truthfulness was less than .30, and inconclusive if the probability fell between the two cutoffs. The variable subsets, outcomes, and area under the ROC curve are presented in Table 12. Results for the standardization sample are presented in parentheses.

Typically, standardization samples provide inflated estimates of decision accuracy, whereas validation samples provide unbiased estimates of accuracy. As expected, accuracy generally was higher in the standardization sample than in the validation sample. The subset that achieved the highest accuracy on cross-validation consisted of EDR amplitude, cardiograph phasic increase in baseline, and respiration line length. Excluding inconclusive outcomes, decision accuracy for this discriminant function on cross-validation was 81.5% correct when the suspect was truthful and 87.9% correct when the suspect was deceptive. Standardized discriminant function coefficients and structure coefficients for the optimal subset are presented in Table 13. The structure coefficient for a variable was the correlation between the variable and the discriminant scores. It was a measure of the extent to which the discriminant score depended on the variable. Although the cardiograph measure correlated .61 with discriminant scores, it did not make a significant unique contribution to the discriminant function. The most heavily weighted variable in the discriminant function was electrodermal amplitude. This was followed by respiration line length and cardiograph phasic increase in baseline.

Table 12. Proportions of Variance Explained (R2), Percent Decision Outcomes and Area under ROC Curve for Validation (and Standardization) Samples for Three Discriminant Functions

		Truthful Cases (n=40/sample)			Deceptive Ca (n=40/sampl	
Discriminant Function	**R^2**	**Correct**	**Wrong**	**Inconc**	**Correct**	**Wrong**
EDR amplitude Cardiograph phasic increase in baseline Respiration line length	.36 (.49)	55 (73)	13 (5)	33 (23)	70 (75)	10 (10)
EDR amplitude Cardiograph tonic increase in baseline Respiration line length	.35 (.52)	53 (70)	15 (5)	33 (25)	65 (78)	13 (10)
EDR amplitude Cardiograph tonic increase in baseline Respiration line length Respiration blocking	.32 (.52)	55 (68)	18 (3)	28 (30)	65 (80)	13 (5)

Table 13. Structure and Standardized Discriminant Function Coefficients for the Optimal Subset of Three DoDPI Criteria

Physiological Cue	Structure Coefficient	Discriminant Coefficient
Electrodermal amplitude	.93	.82 **
Cardiograph phasic increase in baseline	.61	.15
Respiration line length decrease	.55	.36 *

*p < .05. ** p < .01.

Discussion

Analyses of computer measurements of the DoDPI criteria in the standardization sample of 80 field cases from Dollins et al. (2000) supported the use of the following nine DoDPI criteria: EDR amplitude, cardiograph phasic increase in baseline, cardiograph tonic increase in baseline, cardiograph increase in pulse amplitude, respiration line length decrease, respiration blocking (apnea), respiration phasic baseline increase, respiration tonic baseline increase, and respiration gradual decrease in respiration amplitude. None of the other 14 criteria received any empirical support. Of the nine criteria with significant validity coefficients, only five had internal consistency reliabilities that exceeded .33 and significantly correlated with deceptive status in an independent sample of laboratory subjects (Kircher et al., 2001). Of those five measures, three were significantly correlated with deceptive status in the field validation sample. These measures were electrodermal amplitude, cardiograph phasic increase in baseline, and respiration line length decrease. The same three variables were combined by means of a discriminant function that achieved the greatest level of discrimination between truthful and deceptive cases in the validation sample. The same three variables were identified by Harris et al. (2000) as those that capture all of the diagnostic information available in the DoDPI criteria. The same three variables have been used by the CPS program since its introduction in 1991 (Kircher & Raskin, 2001).

Reliability of DoDPI Criteria

An internal consistency reliability of .33 was low compared to standards of psychological testing (APA, 1999). Reliabilities as low as .33 were not a serious concern for several reasons. First, the estimates were conservative. Although in most cases, there were three or four relevant questions per chart, we limited analyses of reliability to the first two relevant questions to achieve equal numbers of items per case. Since reliability depends on the number of items on a test, higher reliabilities would be expected for cases with more relevant questions. Second, .33 was viewed as a minimum and was chosen to provide some representation of each physiological channel in the variable selection process for the discriminant analysis. The reliabilities for variables actually selected for the discriminant function ranged from .43 (respiration line length) to .68 (EDR amplitude). Finally, internal consistency was measured to cast doubt on unproven measures. A new measure that correlates with the deceptive status but is not reliable across items within the test is suspect. The validity coefficient might be spurious and the result might not generalize to another sample of cases. However, questions about internal consistency are less important for measures that already have proven their worth in laboratory and field studies by different teams of investigators over many years.

Electrodermal Measures

Contrary to predictions, EDR complexity and EDR duration were not diagnostic of deception in the either the standardization or validation samples. These null results are in conflict with those obtained in many other studies (e.g., Honts, Raskin, & Kircher, 1994; Kircher & Raskin, 1988; Kircher et al., 2001; Podlesny & Kircher, 1999; Podlesny & Truslow, 1993). The data for the present study were collected in the field, whereas the data in the earlier studies had been collected in simulated laboratory experiments. Although differences between laboratory and field settings could account for the discrepant findings, the two settings

otherwise yield similar patterns of physiological response (Kircher, Raskin, Honts, & Horowitz, 1994).

It is likely that differences in instrumentation account for the discrepant results. In the earlier studies, skin conductance was recorded with a constant 0.5V circuit and wet electrodes according to standards established by the scientific community (Fowles, Christie, Edelberg, Grings, Lykken, & Venables, 1981). In contrast, the electrodermal data for the standardization and validation samples were collected on Axciton polygraphs. The electrodes for these systems are polarizing dry metal plates. The Axciton filters low frequency components from the raw signal, and it alters the activity of the eccrine sweat glands by passing high levels of current through the sweat ducts. Cestaro (1998) found that the electrodermal signals generated by the Axciton did not accurately reproduce known changes in conductance or resistance. Not only were the Axciton signals inaccurate, they were not even monotonically related to the inputs. Consequently, the null results from the Axciton for these two electrodermal measures are not representative of the effects of deception on electrodermal activity when it is recorded properly. There is ample evidence that EDR duration and to a lesser extent EDR complexity are diagnostic, but if they are to remain in DoDPI's list of scoring criteria, the electrodermal data should be collected according to scientifically acceptable recording techniques.

Cardiograph Measures

Phasic increase in baseline was measured from the lowest point following question onset to the highest point in the scoring window (peak amplitude). Tonic increase in baseline was measured by correlating a step function with the mean of systolic and diastolic points over time. A large measure of tonic increase occurred when the cardiograph baseline showed an abrupt increase at question onset that persisted over the 20-second poststimulus scoring window. Conversely, a large measure of phasic increase occurred regardless of whether a large increase in baseline was temporary (phasic) or permanent (tonic). Phasic increase in baseline as measured in the present study would be more aptly named *increase in baseline*.

Phasic increase in baseline was correlated with deceptive status in both the standardization and validation samples. This finding is consistent with those reported in other studies (Harris, Horner, & McQuarrie, 2000; Honts & Kircher, 1994; Kircher & Raskin, 1988; Kircher et al., 2001; Podlesny & Kircher, 1999; Podlesny & Truslow, 1993). Tonic increase in baseline also was diagnostic in the standardization sample ($r = .48$) and in Kircher et al. (2001; $r = .45$), but it only approached significance in the validation sample ($r = .19$, $p < .10$) and in Podlesny and Kircher (1999; $r = .21$, $p < .06$).

Predictably, tonic increase was highly correlated with phasic increase ($r = .73$) and did not predict deceptive status after controlling for phasic increase. These findings indicate that maximum increase in cardiograph baseline is diagnostic regardless of whether it is temporary or permanent.

Cardiograph tonic decrease in baseline was negatively correlated with deceptive status. This finding is contrary to the Swinford (1999) scoring rules. In the standardization sample, almost two-thirds of the probable-lie questions presented to deceptive suspects immediately followed a relevant question, which is common in the modified general question test format (MGQT). When the subject was deceptive to a relevant question, the cardiograph response had not recovered fully before the probable-lie question was presented. The apparent "response" to the probable-lie was a drop in the cardiograph baseline. In actuality, the observed decrease was the recovery to a large response to the preceding relevant question. Deceptive subjects showed decreases in cardiograph during presentations of probable-lie questions, but those decreases were not independent of large cardiograph responses to the relevant questions that preceded them.

Conversely, in most laboratory experiments, each relevant question is preceded by a probable-lie question and followed by a neutral question (Utah Zone Comparison (ZC) format). Here again there is a relationship between decreases in baseline and deceptive status, but in this case, the baseline decreases (recovers) for innocent subjects during relevant questions after showing large cardiograph responses to the

preceding probable-lie question (Kircher et al., 2001; Podlesny & Kircher, 1999). In the MGQT and the ZC, cardiograph tonic decrease in baseline is diagnostic, but it is an artifact of question sequencing, it depends on the response to the question that preceded it, and it should not be scored in the manner described by Swinford (1999).

The law of initial values (Stern, Ray, & Quigley, 2001) predicts that if the subject shows a strong response to one question and it has not recovered fully before the next question is presented, then the response to the next question will be attenuated. In the Utah ZC, probable-lie questions are always preceded by neutral questions, and relevant questions always are preceded by probable-lie questions. Theoretically, this format should benefit the truthful suspect by systematically disattenuating responses to probable-lie questions and attenuating responses to relevant questions. Truthful suspects have less of an advantage in the MGQT, since probable-lie questions sometimes follow relevant questions. Kircher et al. (1994) and Krapohl and McManus (1999) reported that the PLT is biased against truthful suspects because the difference between probable-lie and relevant questions is less in absolute magnitude for truthful suspects than for deceptive suspects. Therefore, there may be some advantage in using a test structure that counters this bias and favors the truthful suspect. This is especially true if decisions are based on numerical evaluations with symmetric +/-6 cutoffs, which is common.

To minimize the effects of initial level on the magnitude of evoked responses, polygraph examiners should wait until the cardiograph baseline returns to the prestimulus level before the next question is presented. Second-by-second plots of cardiograph baseline suggest that it may take as long as 35 seconds for cardiograph responses to recover following a strong response to a test question. In lieu of waiting, the polygraph examiner may present a neutral question. Although the response to the neutral question may be attenuated, it is not formally evaluated and should not affect the test outcome.

In the standardization sample, cardiograph pulse amplitude increase was diagnostic; pulse amplitude decrease was not. Second-by-second plots of pulse amplitude revealed an initial 10-15% reduction in pulse amplitude followed by a recovery that sometimes rebounded beyond the mean prestimulus amplitude by about 5%. The rebound was greater for deceptive responses to relevant questions than to probable-lie questions. Truthful subjects did not respond differentially to probable-lie and relevant questions.

Small but reliable differences between the groups in pulse amplitude increase were observed in other samples as well (Kircher et al., 2001; Podlesny & Kircher, 1999). Given that changes in pulse amplitude are superimposed on relatively large changes in baseline, the effect on pulse amplitude may be too small for most interpreters to detect reliably from a visual inspection of the charts. In the present study, only 3 of the 32 interpreters used increases in pulse amplitude to assign their numerical scores. If the polygraph data are collected by computer, a high-pass digital filter may be used remove changes in baseline and highlight effects on pulse amplitude (Harris et al., 1999; Kircher & Raskin, 1999). Alternatively, the computer could display a derived signal that continuously tracks and displays changes in pulse amplitude (Krapohl, personal communication, January, 2005). However, since pulse amplitude does not start to recover from its initial decrease until the 5th or 6th poststimulus second and does not reach its maximum until the 13th-15th second, polygraph interpreters should be aware that the rebound effects may not occur until after the subject has answered the test question. If the response begins after the subject has answered, it is too late according to current DoDPI guidelines.

Changes in pulse rate were not significantly correlated with deceptive status in the standardization or validation samples. The null results for decreases in pulse rate were not consistent with the significant decelerations reported elsewhere (e.g., Kircher et al., 2001; Patrick & Iacono, 1991; Podlesny & Kircher, 1999; Podlesny & Truslow, 1993).

Respiration Measures

Computer measurements of DoDPI criteria for thoracic and abdominal respiration

channels were correlated (r = .58). Numerical evaluations of thoracic and abdominal respiration also were correlated (M_r = .92). Given the strength of these correlations and the long-standing difficulties of deriving reliable measures of differential reactivity from respiration recordings, it is advisable to combine these measures to improve reliability (Nunnally, 1978). This recommendation is consistent with current practice in both human (Bell et al., 1999; Swinford, 1999) and computer analysis of respiration recordings (Kircher & Raskin, 2001).

Apnea and line length were more diagnostic and more reliable than measures of respiration amplitude, rate, or baseline. Each measurement of apnea was based on 40 samples of respiration (250 ms samples for 10 seconds following question onset), and each measurement of line length was based on 600 samples (10 seconds at 60 Hz). In contrast, measurements of amplitude, rate, I/E ratio, and baseline were based on only six cycles and were much less reliable. When the reliability of a measure is low, the magnitude of the effect has to be large for the measure to be useful.

Of 10 respiration criteria that were measured over two or three cycles, four yielded at least one combined respiration measure that was significantly correlated with deceptive status. Three of the four significant validity coefficients were associated with three-cycle models. Two other models predicted changes that persisted over four cycles (respiration baseline tonic increase and respiration baseline tonic decrease), both of which were significantly correlated with deceptive status. Thus, when changes in respiration were diagnostic, they generally lasted 3 or 4 cycles. Only 1 of 10 2-cycle changes was diagnostic and it had a reliability coefficient of only .17. These findings suggest that only 3- or 4-cycle changes in respiration are large enough to be diagnostic.

Principal components analysis revealed four sources of variance among computer measures of the various respiration criteria. The components of respiration measures were rate, amplitude, baseline, and apnea. As expected, the most reliable and valid respiration measure was line length. Line length loaded on apnea, amplitude, and rate components, and it captured the diagnostic variance in all of those components. Changes in respiration baseline also were diagnostic and contributed significantly to the regression equation for predicting deceptive status from combined respiration cues.

In the standardization sample, respiration line length and respiration baseline were largely independent sources of diagnostic variance. There has been considerable empirical support for the use of respiration line length for the detection of deception since Timm introduced this measure in 1982 (Harris et al., 2000; Honts, Raskin, & Kircher, 1994; Horowitz, Kircher, Honts, & Raskin, 1997; Kircher et al., 2001; Kircher & Raskin, 1988; Kircher et al., 1994; Podlesny & Kircher, 1999; Timm, 1982). Significant effects for respiration phasic and tonic baseline increase were reported by Raskin and Hare (1978) and were found in a sample of confirmed criminal cases (Raskin, Kircher, Honts, & Horowitz, 1988). Diagnostic decreases in respiration baseline also were reported by Raskin and Hare (1978) but were not evident in the field study by Raskin et al. (1988). No diagnostic increase or decrease in respiration baseline was observed in the present validation sample or in several other studies (Kircher et al., 2001; Patrick & Iacono, 1991; Podlesny & Kircher, 1999). Thus, effects of deception on respiration baseline have been reported previously, but they generally are small and appear in only some data sets. There is more support for the continued use of baseline increase than baseline decrease. If respiration baseline decrease is to be retained as a DoDPI criterion, its occurrence should be viewed as an indication of relief rather than arousal or concern.

Lens Models

Correlations between interpreters' numerical evaluations and computer measurements of the various DoDPI criteria (cue utilizations) generally followed cue validities. That is, interpreters appeared to use a cue to assign their numerical scores to the extent that the cue was diagnostic. Not only did interpreters use valid DoDPI criteria, they also avoided using invalid criteria. Apparently, these interpreters, many of whom were instructors at DoDPI, knew from experience not to use some of the DoDPI criteria.

That interpreters made optimal use of the physiological cues was evident from the matching coefficients (*G*). The matching coefficient summarized the agreement between the optimal regression weights for predicting deceptive status and the regression weights for predicting the interpreter's numerical evaluations. Generally, interpreters made optimal use of the electrodermal (M_G = .99) and cardiovascular criteria (M_G = .97) and suboptimal use of respiration criteria (M_G = .83).

The *C* coefficient from the lens model was large to the extent that the interpreter was correct about the individual's deceptive status when the computer was wrong. The *C* coefficients for 15% to 25% of the interpreters in the present study were statistically significant. Even when the *C* coefficients were significant, they rarely exceeded .30. The largest values of *C* were obtained for the cardiograph. This suggests that if more diagnostic information is to be extracted by the computer from physiological recordings, such efforts should focus on the cardiograph.

The *C* coefficients in the present study generally were smaller than those reported by Kircher et al. (1995). Kircher et al. did not conduct separate lens model analyses of the numerical scores for different channels. Rather, they analyzed the total numerical score across all channels and were not specific about the source of the large *C* coefficients for their interpreters. In addition, only paper charts were available to Kircher et al. The data were entered into the computer by manually tracing each of the signals on a digital tablet. In many cases, the skin resistance recording had hit a pen stop and was truncated. The signals were entered into the computer exactly as they appeared on the paper chart, and no attempt was made to predict how large the EDR would have been it if had not hit the pen stop. Since the human evaluator was free to make such predictions while scoring the charts and the computer was not, the computer was disadvantaged in its analysis of the most diagnostic physiological measure. The large *C* coefficients in Kircher et al. may have been due to interpreters' abilities to infer EDR amplitude from the truncated electrodermal signals. In the present study, the computer and the polygraph interpreters had access to the same data, and the *C* coefficients were not as large as those reported by Kircher et al. (1995).

Lens model analyses of the electrodermal data in the present study indicated that numerical scores for the electrodermal channel were based exclusively on EDR amplitude for all but two of the interpreters. This strategy was appropriate since only EDR amplitude was predictive of deceptive status.

Lens model results were less clear for the cardiograph data. About half of the interpreters had significant regression coefficients for cardiograph phasic increase in baseline. The other half had significant regression coefficients for tonic increase in baseline. However, since these two measures were highly correlated, one measure could be dropped from the regression equation and the other would account for the variance in the numerical scores. Only 13% of the interpreters used cardiograph tonic decrease in baseline. Interpreters who used decrease in baseline did not treat it as an indication of arousal, as recommended by Swinford (1999). Rather, they treated decrease in baseline as an indication that the subject was relatively unconcerned about the question. The interpreters were correct; the criterion was wrong.

The agreement between interpreters' models for predicting numerical scores and the optimal combination of physiological cues was lower for respiration (M_G = .83) than for the other two channels. Computer analysis of all three physiological measures was more accurate than that of the average human interpreter, but the difference was greatest for respiration. Whereas computer measures of respiration accounted for 28% of the variance in deceptive status, numerical scores, on average, accounted for only 9% of the variance. Cue validity for the computer model was almost two standard deviations greater than the mean for the 32 polygraph interpreters, and not one interpreter performed as well as the computer in scoring respiration.

Respiration line length and tonic increase in baseline were significant predictors of deceptive status in the computer model of respiration cues. Although 88% of the interpreters relied on line length, only 16% relied on baseline increase. These results suggest that numerical evaluations of

respiration would have been more accurate if the interpreters had attended more to increases in respiration baseline. However, any general recommendation to attend more to respiration baseline must be tempered by the fact that this DoDPI criterion was diagnostic in only about half of the studies in which it was evaluated. On the other hand, for most interpreters, reliance on computer or other reliable measurements of line length would significantly increase the accuracy of their numerical evaluations of respiration.

The results of the principal components analysis suggest that observed changes in apnea, amplitude, and rate could be used *in combination* to estimate line length if actual measurements of line length are unavailable. However, we suspect that DoDPI-trained interpreters already use this approach to assign numerical scores. If so, their performance in this study suggests that this approach is ineffective, and a computer or other objective measurement of line length should be used in lieu of human judgments.

One of six individual difference variables was related to the accuracy of numerical evaluations. Years of experience correlated .40 with the accuracy of numerical evaluations of electrodermal responses. The most experienced interpreters may have automated the basic skills required to score electrodermal responses. If so, they would have freed up working memory that then could be used to consider other aspects of the recordings as they assigned numerical scores. Other aspects would include deep breaths or countermeasure maneuvers. It is also possible that the finding is spurious. As compared to the electrodermal channel, there was more variance among interpreters in how they used the respiration and cardiograph criteria. As compared to the electrodermal channel, there was more opportunity to show the benefits of experience in scoring respiration or the cardiograph, and yet, neither of those relationships materialized. Having too many rules for scoring those channels or having rules that are not well articulated may have contributed to the variance among interpreters.

The accuracies of numerical scores assigned to the cardiograph and to respiration correlated .38. Interpreters who extracted more useful information from the cardiograph also tended to extract more useful information from respiration recordings. Expertise in scoring these channels might be used to select interpreters for quality control, education, and training programs or to inform further development of automated scoring systems.

Computer Model of DoDPI Criteria

Over 150 different computer measurements were extracted from the standardization sample to characterize the 23 DoDPI criteria described by Swinford (1999). Most of the variables represented slightly different ways of measuring the same type of physiological response and were largely redundant. For example, respiration phasic increase in baseline was measured for thoracic, abdominal, and combined respiretion channels. It was measured relative to prestimulus baseline and again relative to the chart mean baseline value. Finally, it was measured over two and three poststimulus cycles. With many variables, it is a simple matter to develop a statistical classifier that will distinguish between two groups of subjects with perfect or near perfect accuracy. It is far more challenging to develop a computer model for discriminating between *populations* of truthful and deceptive subjects (Kircher & Raskin, 2001).

To achieve this objective, we identified a small number of relatively independent sources of diagnostic information in the polygraph charts. Variables were retained with internal consistency reliabilities that exceeded .33 and had significant correlations with deceptive status in the standardization sample. Several archival data sets were resurrected and reconditioned to assess the generalizability of results from the standardization sample before discriminant functions were developed and tested on the validation sample. The only archival data set formally included in the present report consisted of the 84 subjects in the probable-lie neutral and effective feedback conditions from Kircher et al. (2001). However, additional data sets were used on an ad hoc basis to study the dependability of results from the standardization sample. These included 80 subjects from Podlesny and Kircher (1999), 63 criminal suspects from Raskin et al. (1988), and 60 subjects from Kircher and Raskin (1988).

Of five variables that survived the selection process, three were significantly correlated with deceptive status in the validation sample: EDR amplitude, cardiograph phasic increase in baseline, and respiration line length. The same three variables were selected for one of the three discriminant functions that were tested on the validation sample. They also comprise the set of variables that has been used by CPS for many years (Kircher & Raskin, 2001).

The three discriminant functions performed similarly on cross-validation. They varied in terms of the area under the ROC curve by only 1%. The largest ROC value on cross-validation was .856 and it was achieved by the CPS subset. However, the weights for the variables that were optimal for the standardization sample differed somewhat from those used by CPS. Both the CPS model and the standardization model weighed EDR amplitude more heavily than cardiograph increase in baseline or respiration line length. However, as compared to CPS, the standardization model gave more weight to EDR amplitude and less weight to cardiograph increase in baseline. Interestingly, if the discriminant function had been based on the validation sample, as compared to CPS, *less* weight would have been given to EDR amplitude and *more* weight would have been given to the cardiograph. These findings suggest that the CPS weights would have been optimal if the standardization and validation samples had been combined.

Excluding 26.5% inconclusive outcomes, decisions by the standardization model on cross-validation were 80.9% correct on truthful cases and 87.5% correct on deceptive cases. As discussed elsewhere (Kircher & Raskin, 2000, 2001), the CPS variable subset is robust. It consistently achieves respectable levels of decision accuracy in laboratory and field settings, even with substandard recording equipment. However, in 20-25% of cases, it requires more than three repetitions of the question sequence (charts) to reach a definite decision. The present study was limited to the first three charts of recorded physiological data. Had more charts been available for cases that were inconclusive after three charts, we would expect the inconclusive rate to drop to between 10% and 15% and little change in the accuracy of decisions.

Limitations and Areas for Additional Research

In the present study, the term *cue validity* was used to characterize the relationship between a physiological measure or set of measures and deceptive status. The observed relationships in our samples provided estimates of the diagnosticity of these measures in the field. Questions always may be raised about the extent to which the samples of truthful and deceptive cases in our studies adequately represent the target populations of truthful and deceptive suspects (e.g., Iacono & Lykken, 1997; Krapohl, Shull, & Ryan, 2002; Raskin, Honts, & Kircher, 1997). The procedures for selecting cases from archives of field polygraph examinations are described above and in Dollins et al. (1999), and characteristics of the standardization and validation samples are presented in the Table 1. In our judgment, the procedures used by Dollins et al. were adequate to represent the population of polygraph examinations conducted by federally trained examiners, and our results are representative of the accuracies achieved in criminal investigations by federally-trained field polygraph examiners using Axciton polygraphs. Others may disagree.

Interpreters were asked to use DoDPI criteria to score the charts. The criteria that were current at the time the interpreters scored the charts (Stern, 2003) were not the same as those measured by the computer and included in the lens model analysis (Swinford, 1999). Specifically, three of the DoDPI criteria included in the present study had been excluded from the 2003 DoDPI Test Analysis manual. The three omitted criteria were cardiograph decrease in baseline, apnea (holding), and respiration decrease in baseline. The 2003 criteria were an improvement over the 1999 criteria because the omitted criteria were either uncorrelated (holding) with deceptive status, or they were significantly related to deceptive status, but the relationship was opposite to the scoring rules described by Swinford (1999; cardiograph tonic decrease in baseline and respiration decrease in baseline).

In the field, polygraph examiners interact with the subject and often have intimate knowledge of the case. Numerical scores should be based on only on the

polygraph charts, but extrapolygraphic information could affect the scores assigned by the original examiner. In the present study, there was little evidence that the original examiner benefited from extra-polygraphic information because the accuracy they achieved in the standardization sample (ROC = .908) was similar to that achieved by study participants (Mean ROC = .882). However, other data suggest that decisions by the original examiner often are more accurate than those by independent numerical scorers (Raskin et al., 1997). Additional research is needed to determine if extrapolygraphic in-formation benefits the original examiner; and if so, can that information be quantified and combined with the results of the polygraph examination in a formal way to reach a decision that is optimal for the particular testing context (Kircher & Raskin, 2001).

After the interpreters had completed all of their numerical evaluations, a computerized sampling strategy was used to extract certain sections of polygraph charts they had scored previously. These sections subsequently were projected from the computer to a large screen. Interpreters were interviewed individually. They were given a laser pointer and asked to verbalize their thoughts while they scored the chart segment. The interview was videotaped. These data may provide some valuable insights into interpreters' decision processes. Of particular interest are over-sampled occasions where the interpreter disagreed with the computer and the interpreter was correct. From these interviews, it might be possible to develop more valid computer algorithms for scoring charts.

Prior to receiving individualized feedback about their performance, interpreters who were not interviewed were presented with all possible pairings of the DoDPI criteria and were asked to choose the most useful from each pair. In the future, these data may be used to scale the DoDPI criteria in terms of perceived utility and compare those ratings to measures of actual use obtained from the lens models.

Finally, the lens model might be used to assess students' ability to evaluate polygraph charts. Periodically during training, students would score a set of charts. Lens model analyses of their numerical evaluations would reveal their particular strengths and weaknesses in scoring charts. If problems were uncovered, corrective measures could be taken to improve students' interpretive skills before they complete their training.

References

American Psychological Association (APA) (1999). Standards for Educational and Psychological Testing. American Psychological Association: Washington, D.C.

Ansley, N. & Krapohl, D, J. (2000). The frequency and appearance of evaluative criteria in field polygraph charts. *Polygraph*, 29(2),

Backster, C. (1962). The Backster chart and reliability rating method. *Law and Order*, 11, 63-64.

Backster, C. (1969). *Technique fundamentals of the tri-zone polygraph test.* New York: Backster Research Foundation.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12, 387-415.

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

Ben-Shakhar, & E., Lielich, I. (1984). On statistical detection of deception: Comment on Szucko and Kleinmuntz. *American Psychologist*, 39, 79-80.

Brunswik, E. (1952). *Conceptual framework of psychology*. Chicago: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of experiments* (2nd Ed.) Berkeley: University of California Press.

Cestaro, V. L. (1998). Instrumentation for presenting a known standard signal to the electrodermal channel for assessing response characteristics. *Polygraph*, 27(3), 188-209.

Fowles, D. C., Christie, M. J., Edelberg, R., Grings, W.W., Lykken, D. T., & Venables, P. H. (1981). Committee report, publication recommendations for electrodermal measurements. *Psychophysiology*, 18(3), 232-239.

Harris, J. C. & Horner, A. (1999). Reformat.Version 1.1. Johns Hopkins University, Applied Physics Laboratory.

Harris, J.C., Horner, A., & McQuarrie, D.R. (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-00-7272.

Honts, C. R., Raskin, D. C. & Kircher, J. C. (1994). Effects of physical and mental countermeasures on the detection of deception. *Journal of Applied Psychology*, 79, 252-259.

Horowitz, S. W., Kircher, J. C., Honts, C. R., & Raskin, D. C. (1997). The role of control questions in physiological detection of deception. *Psychophysiology*, 34, 108-115.

Iacono, W. G. & Lykken, D. T. (1997). The scientific status of research on polygraph techniques: The case against polygraph tests. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), *The Scientific Evidence Manual* (Volume 1) 1997. St. Paul: West Publishing Co.

Kircher, J. C., Horowitz, S. W. & Raskin, D. C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior*, 12, 79-90.

Kircher, J. C. & Raskin, D. C. (1982, October). Cross-validation of a computerized diagnostic procedure for detection of deception. Paper presented at the meeting of the Society for Psychophysiological Research, Minneapolis, MN.

Kircher, J. C. & Raskin, D. C. (1983). Clinical versus statistical lie detection revisited: Through a lens sharply. *Psychophysiology*, 20, 452. (Abstract).

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

Kircher, J. C., & Raskin, D. C. (1999). *The Computerized Polygraph System (Version 3.0) Manual.* Scientific Assessment Technologies, Salt Lake City, UT.

Kircher, J. C., & Raskin, D. C. (2000). CPS Comments on Dollins et al's computer algorithm comparison. *Polygraph*, 29(3), 250-252.

Kircher, J. C. & Raskin, D. C. (2001). Computer methods for the psychophysiological detection of deception. In M. Kleiner (Ed.) *Handbook of Polygraphy*. London: Academic Press.

Kircher, J. C., Raskin, D. C., Honts, C. R., & Horowitz, S. W. (1994). Genereralizability of statistical classifiers for the detection of deception. *Psychophysiology*, 31, S73. (Abstract)

Kircher, J. C., Raskin, D. C., Honts, C. R., & Horowitz, S. W. (1995). Lens model analysis of decision making by field polygraph examiners. *Psychophysiology*, 32, S45. (Abstract)

Kleinmuntz, B. (1963). MMPI decision rules for the identification of college maladjustment: A digital computer approach. *Psychological Monographs*, 77, Whole No. 477.

Kleinmuntz, B. J. & Szucko, J. J. (1982). On the fallibility of lie detection. *Law & Society Review*, 17(1), 85-104.

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

Krapohl, D. J., Shull, K. W., & Ryan, A. A. (2002). Does the confession criterion in case selection inflate polygraph accuracy estimates? *Forensic Science Communication*, 4(3), 1-12.

Kubis, J. F. (1973). Analysis of polygraph data. Part 2. *Polygraph*, 2, 89-107.

Lockette, E. & Kircher, J. C., (2003) Growth curve analysis of polygraph data. (Grant No. DASW01-03-1-0001). Final report to the U.S. Department of Defense. Salt Lake City: University of Utah, Department of Educational Psychology.

McNemar, Q. (1968). *Psychological Statistics* (2nd Ed). New York: Wiley.

Olsen, D. E., Harris, J. C., Capps, M. H., & Ansley, N. (1997). Computerized polygraph scoring system. *Journal of Forensic Sciences*, 42, 61-70.

Patrick, C.J., & Iacono, W.G. (1991). A comparison of field and laboratory polygraphs in the detection of deception. *Psychophysiology*, 28(6), 632-638.

Podlesny, J. A. & Kircher, J. C. (1999). The Finapres (volume clamp) recording method in psychophysiological detection of deception examinations: Experimental comparison with the cardiograph method. *Forensic Science Communication*, 1(3), 1-17.

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

Podlesny, J.A., Truslow, C.M. (1993). Validity of an expanded-issue (Modified General Question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology*, 78(5). 788-797.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis* (2nd Ed.). Thousand Oaks: Sage.

Raskin, D. C. (1976). Reliability of chart interpretation and sources of errors in polygraph examinations. Report to the National Institute of Law Enforcement and Criminal Justice (Contract 75-NI-99-0001). Salt Lake City: University of Utah, Department of Psychology.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (1997). The case for the admissibility of the results of polygraph examinations. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), *The scientific evidence manual* (Volume 1) 1997. St. Paul: West Publishing Co.

Reid, J. E. & Inbau, F. E. (1977). *Truth and deception: The polygraph ("lie detector") technique*. Baltimore: Williams & Wilkins.

Senter, S. M., Dollins, A. B., & Krapohl, D. J. (1999). Comparison of Utah and DoDPI scoring accuracy: Equating scoring rule, chart rule, and number of data channels used. Department of Defense Polygraph Institute Research Division (DoDPI 97-P-0005).

Stern, B. (2003). Test Data Analysis: DoDPI Numerical Evaluation Scoring System. Forensic Psychophysiology Program, Department of Defense Polygraph Institute.

Stern, R. M, Ray, W. J. & Quigley, K. S. (2002). *Psychophysiological recording* (2nd Ed.). Oxford University Press: New York.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph*, 28, 10-27.

Szucko, J. J. & Kleinmuntz, B. J. (1981). Statistical versus clinical lie detection. *American Psychologist*, 36, 488-496.

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

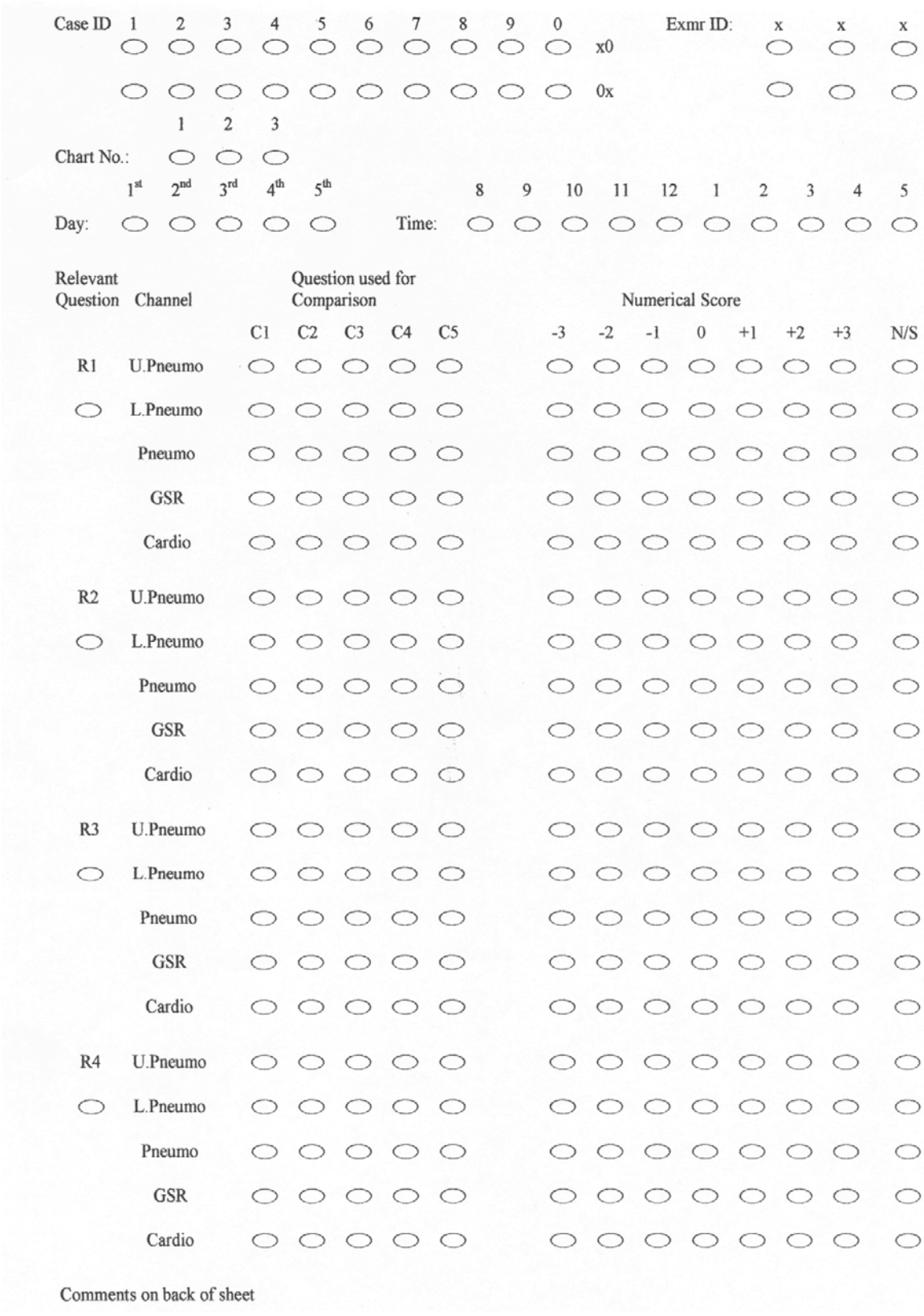
Weaver, R. S. (1985). Effects of differing numerical chart evaluation systems on polygraph examination results. *Polygraph*, 14, 34-41.

Appendix A
Interpreter Demographic Information

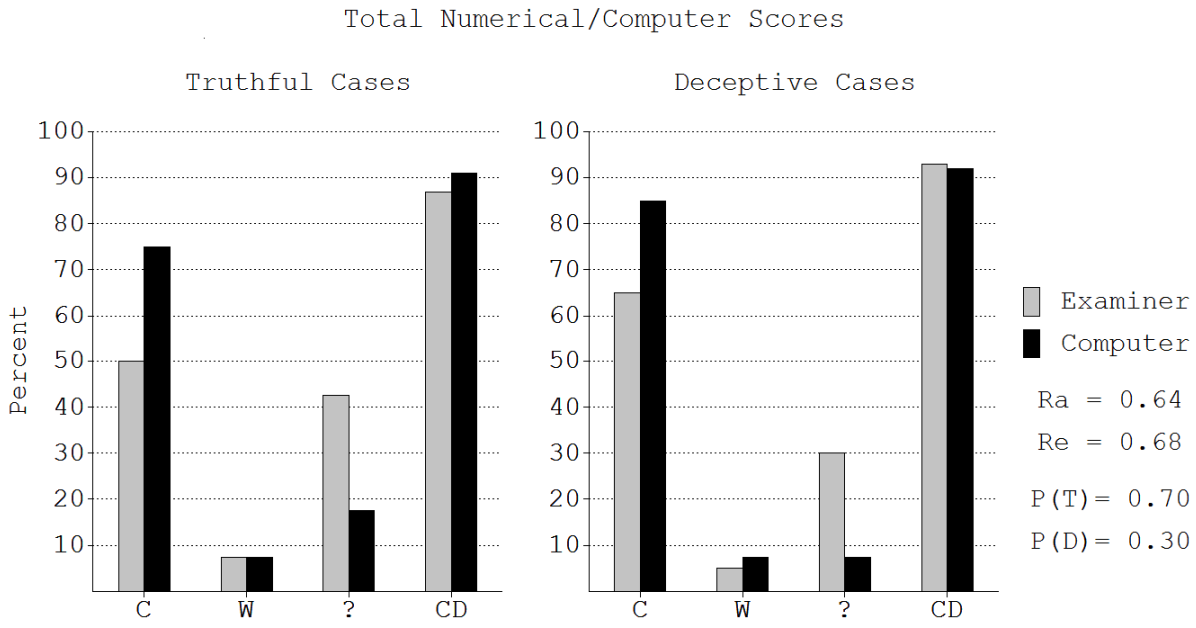
Table 14. Interpreter Demographic Information

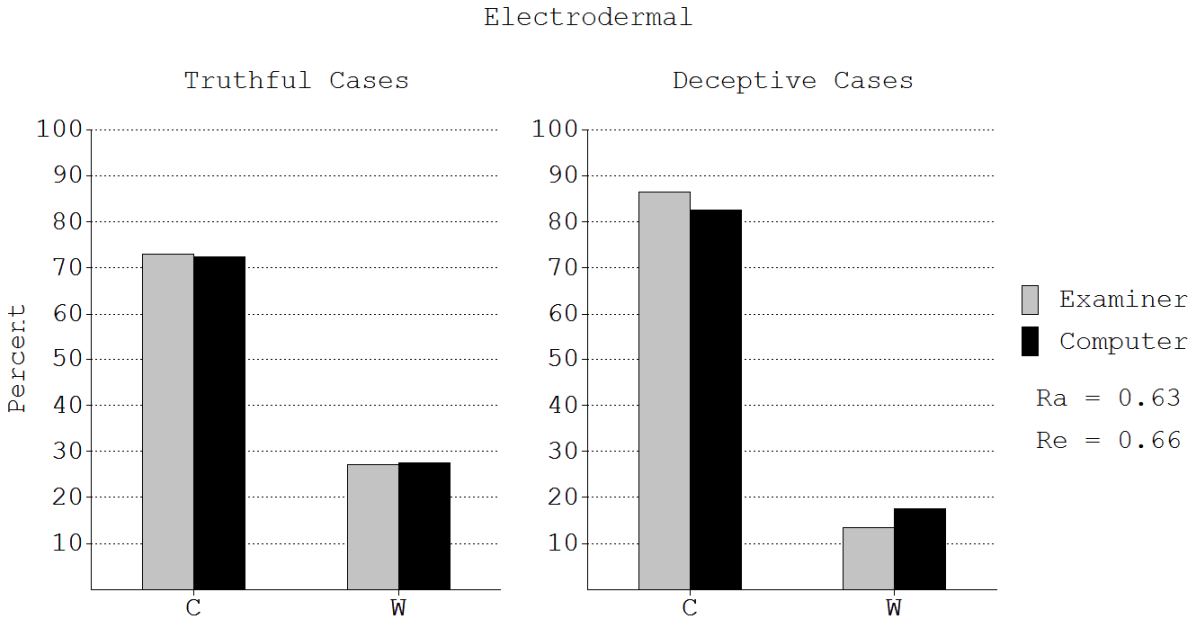
	n	Mean	SD	Min.	Max.	Md
Age (years)	32	45.7	7.5	27	58	45.5
Time in current position (months)	31	60.1	57.4	1	216	42.0
Time evaluating polygraph charts (months)	32	162.9	77.2	26	281	162.0
Total CE hours per yr.	32	50.5	19.0	2	104	40.0
Number of classes taught during career	32	106.6	202.3	0	1000	20.0
Number of CQ tests conducted in the last yr.	32	93.7	193.0	0	1000	30.0
Number of RI tests conducted in the last yr.	32	23.6	57.7	0	250	0.0
Number of GK tests conducted in the last yr.	32	0.1	0.4	0	2	0.0

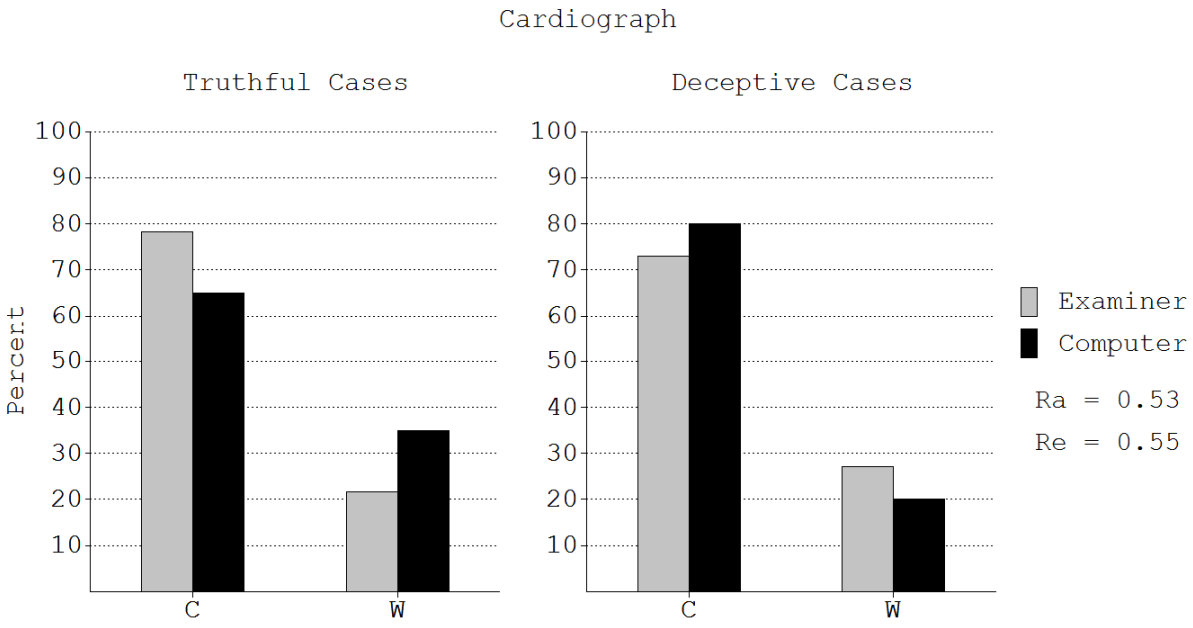
Appendix B
Interpreter Scoresheet

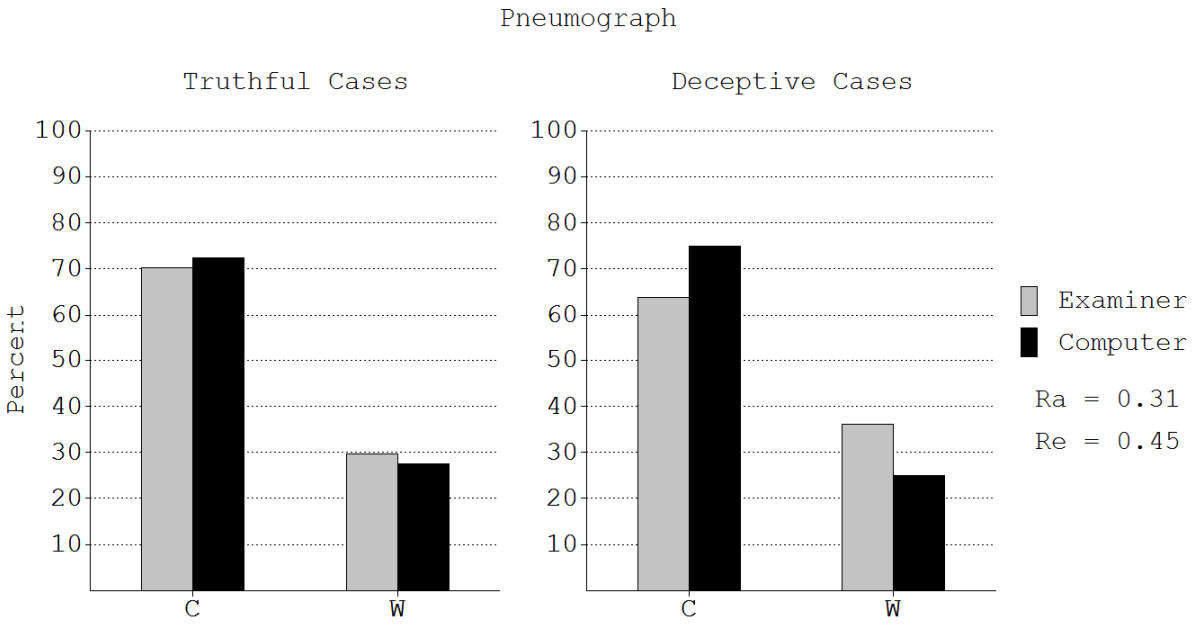


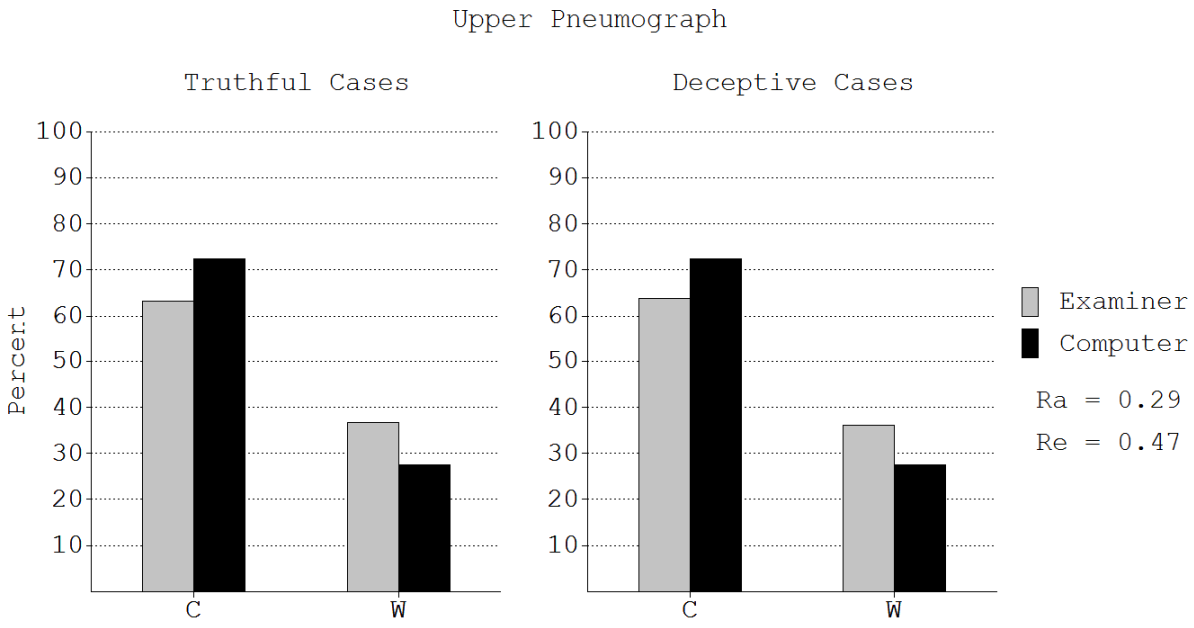
Appendix C
Example of Interpreter Feedback

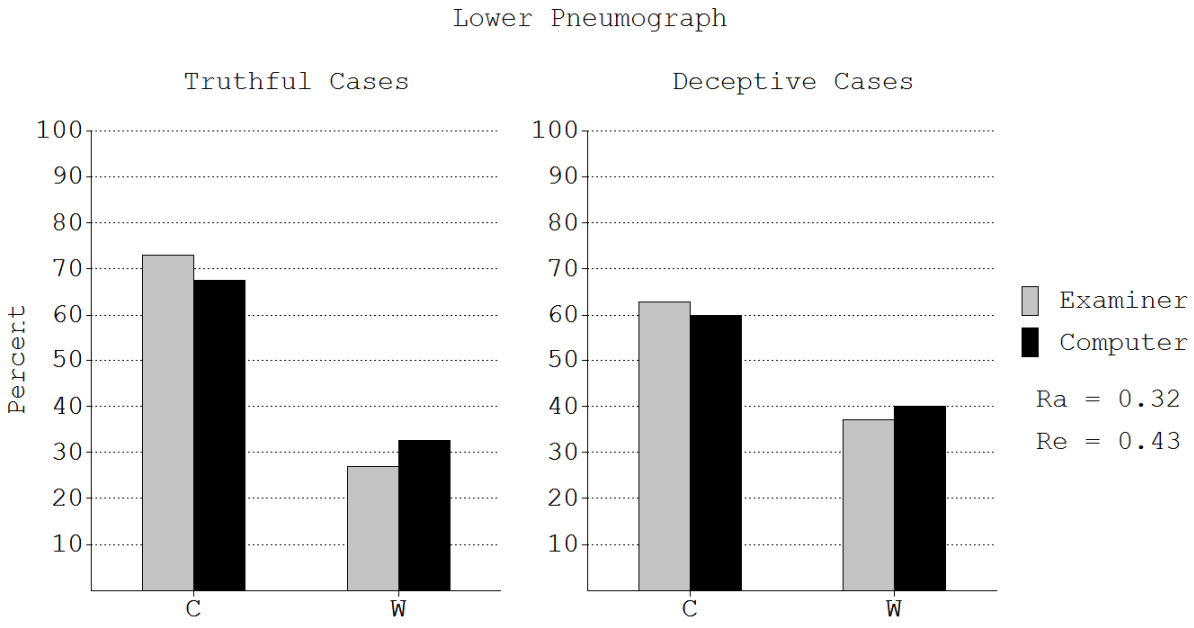


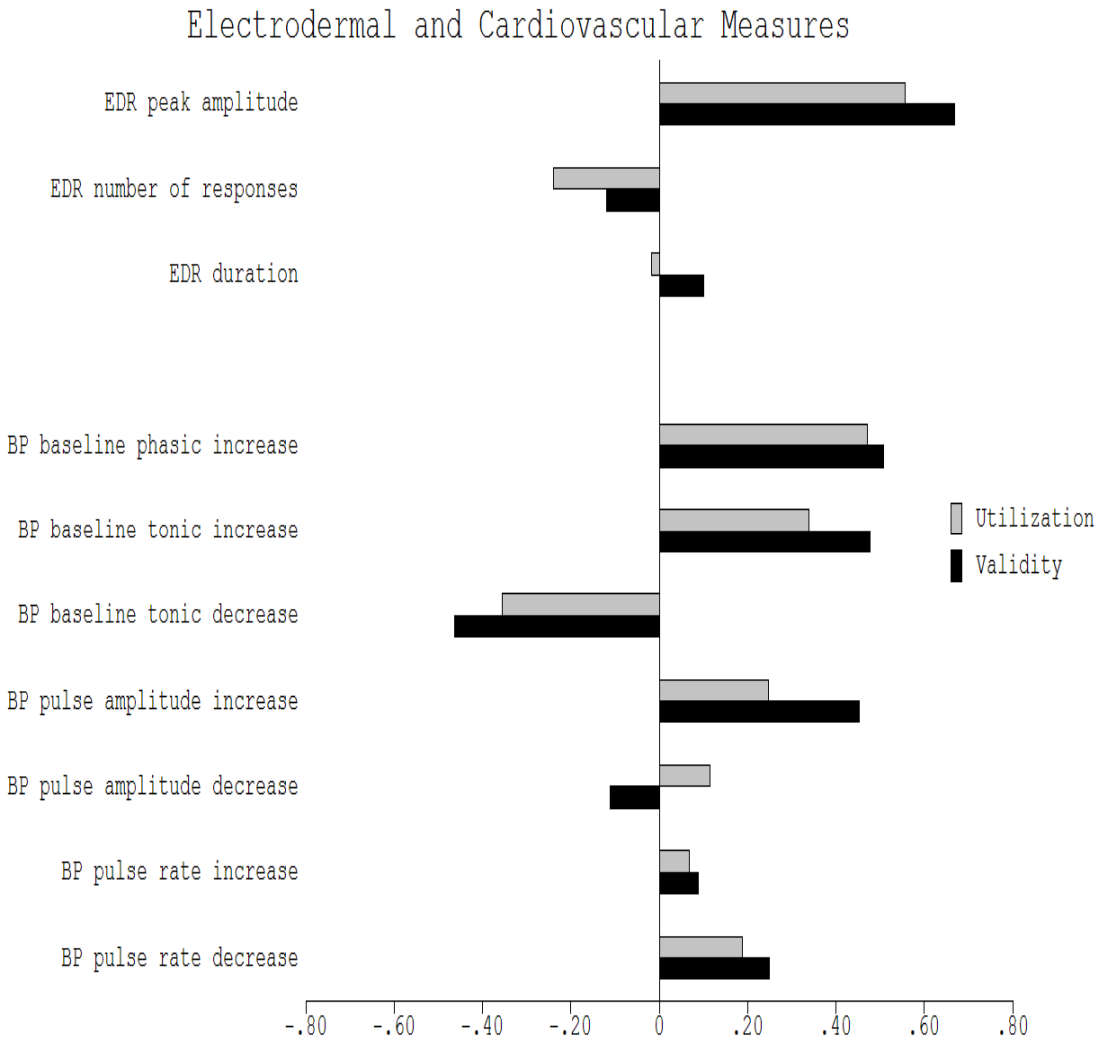


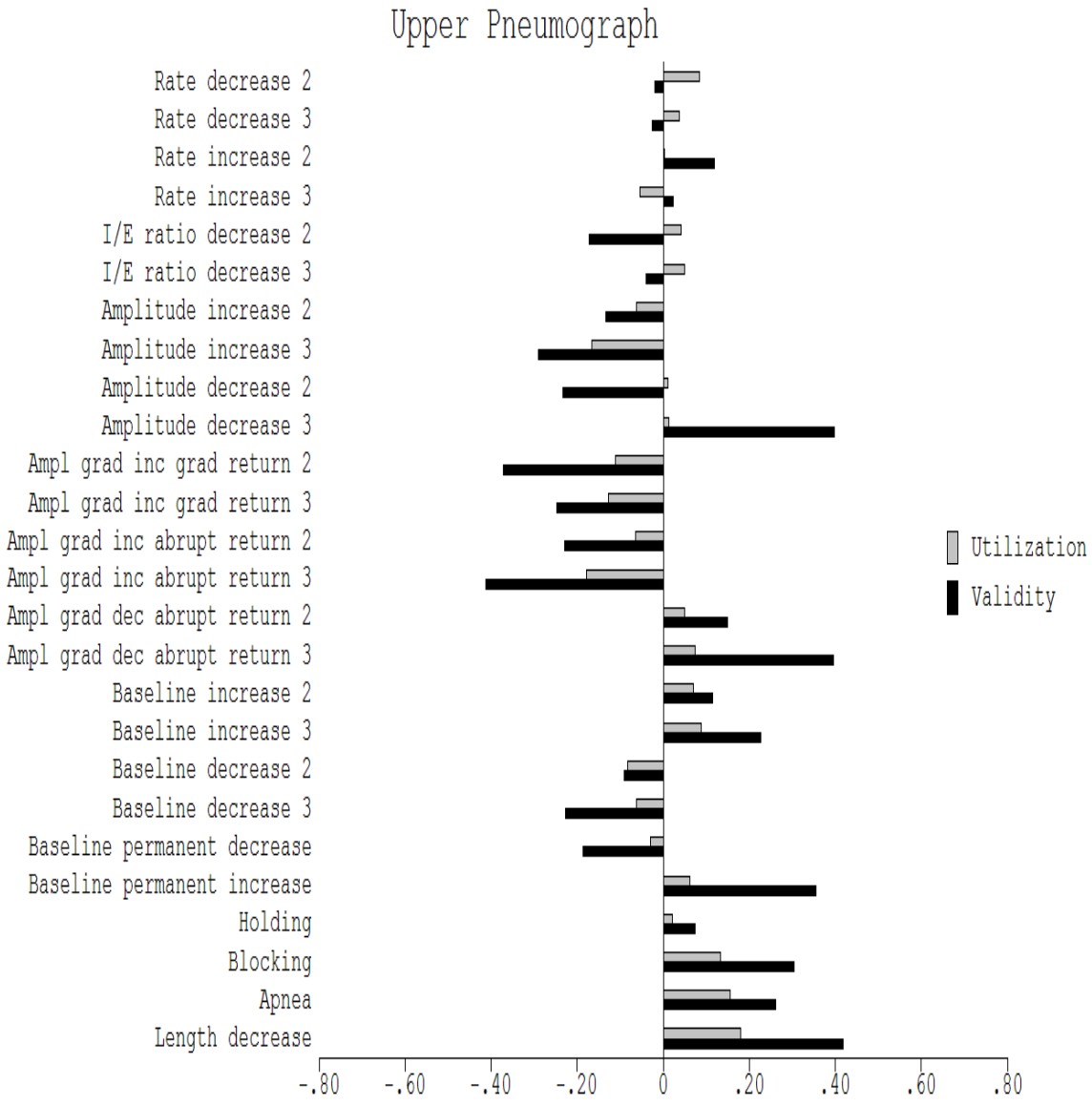

Cardiograph
Truthful Cases
Deceptive Cases
Percent
100
90
80
70
60
50
40
30
20
10
C
W
Examiner
Computer
Ra = 0.53
Re = 0.55

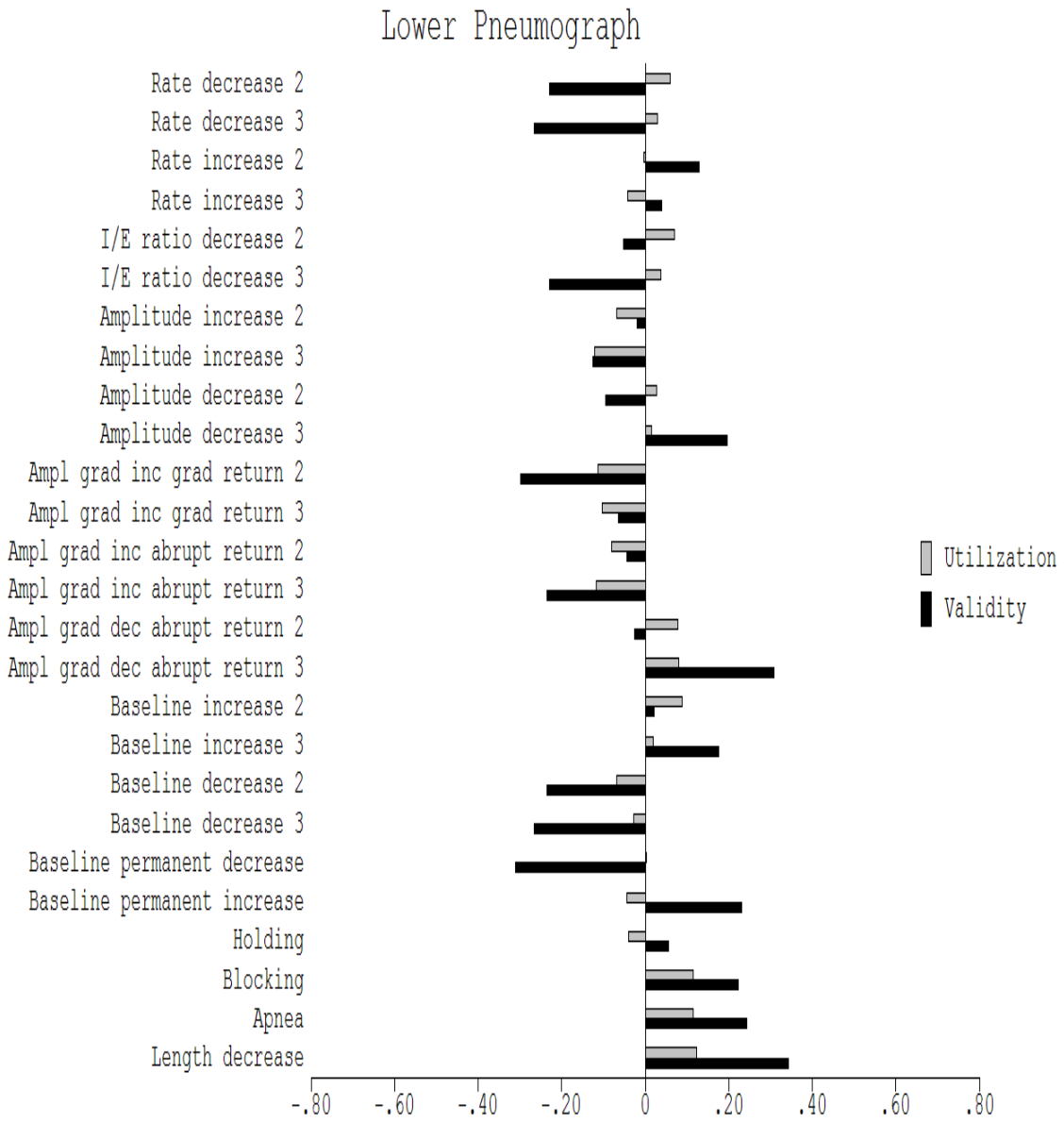

Pneumograph
Truthful Cases
Deceptive Cases
Percent
100
90
80
70
60
50
40
30
20
10
C
W
Examiner
Computer
Ra = 0.31
Re = 0.45


Upper Pneumograph
Truthful Cases
Deceptive Cases
Percent
100
90
80
70
60
50
40
30
20
10
C
W
Examiner
Computer
Ra = 0.29
Re = 0.47


Lower Pneumograph
Truthful Cases
Deceptive Cases
Percent
100
90
80
70
60
50
40
30
20
10
C
W
Examiner
Computer
Ra = 0.32
Re = 0.43


Electrodermal and Cardiovascular Measures
EDR peak amplitude
EDR number of responses
EDR duration
BP baseline phasic increase
BP baseline tonic increase
BP baseline tonic decrease
BP pulse amplitude increase
BP pulse amplitude decrease
BP pulse rate increase
BP pulse rate decrease
Utilization
Validity
-.80
-.60
-.40
-.20
0
.20
.40
.60
.80


Upper Pneumograph
Rate decrease 2
Rate decrease 3
Rate increase 2
Rate increase 3
I/E ratio decrease 2
I/E ratio decrease 3
Amplitude increase 2
Amplitude increase 3
Amplitude decrease 2
Amplitude decrease 3
Ampl grad inc grad return 2
Ampl grad inc grad return 3
Ampl grad inc abrupt return 2
Ampl grad inc abrupt return 3
Ampl grad dec abrupt return 2
Ampl grad dec abrupt return 3
Baseline increase 2
Baseline increase 3
Baseline decrease 2
Baseline decrease 3
Baseline permanent decrease
Baseline permanent increase
Holding
Blocking
Apnea
Length decrease
-.80
-.60
-.40
-.20
0
.20
.40
.60
.80
Utilization
Validity


Lower Pneumograph
Rate decrease 2
Rate decrease 3
Rate increase 2
Rate increase 3
I/E ratio decrease 2
I/E ratio decrease 3
Amplitude increase 2
Amplitude increase 3
Amplitude decrease 2
Amplitude decrease 3
Ampl grad inc grad return 2
Ampl grad inc grad return 3
Ampl grad inc abrupt return 2
Ampl grad inc abrupt return 3
Ampl grad dec abrupt return 2
Ampl grad dec abrupt return 3
Baseline increase 2
Baseline increase 3
Baseline decrease 2
Baseline decrease 3
Baseline permanent decrease
Baseline permanent increase
Holding
Blocking
Apnea
Length decrease
-.80
-.60
-.40
-.20
0
.20
.40
.60
.80
Utilization
Validity