JOHNS HOPKINS UNIVERSITY
Applied Physics Laboratory

11100 Johns Hopkins Road, Laurel, MD 20723-6099

The Preliminary Credibility Assessment System Embedded Algorithm Description and Validation Results

GED-R-06-7571

John C. Harris
Allan D. McQuarrie

Prepared under contract MDA904-02-D-0302, TTO 0034 for:
Counterintelligence Field Activity
251 18th Street, Suite 1200
Arlington, VA 22202

Executive Summary

This report describes a classification algorithm developed by the Johns Hopkins University Applied Physics Laboratory for use in evaluating data collected by the Preliminary Credibility Assessment System (PCASS). The PCASS instrument measures two physiological signals during a specifically generated structured interview and the embedded classification algorithm provides an indication of whether the physiological changes occurring during the interview are associated with truthfulness or deceptiveness. The PCASS is intended for use in an environment where the risk associated with indicating a deceptive interviewee as truthful (false negative) is judged to be significantly more consequential than indicating a truthful interviewee as deceptive (false positive) and this has been taken into consideration when establishing outcome decision rules. The PCASS algorithm provides one of three color-coded outcomes: red, green, or yellow, which respectively indicate changes associated with deceptiveness, truthfulness, or otherwise, inconclusiveness.

The PCASS instrument was developed concurrently with that of the embedded algorithm. Since no data were available from the PCASS itself, a database of similar polygraph data was used to train, test, and validate the algorithm. The combined test and validation datasets consisted of 258 confirmed deceptive and 64 confirmed truthful field polygraph examinations. On the combined test and validation datasets the algorithm evaluated 8% of truthful interviewees as red and only 2% of deceptive interviewees as green. When used to infer truthfulness or deceptiveness from outcome color and assuming a 50-50 base rate of deceptive and truthful interviewees, the algorithm results in conditional outcome accuracies of:

Probability (Deceptive | Red) = 92% and
Probability (Truthful | Green) = 97% with
a 27% overall inconclusive rate (Yellow).

These results may vary when the algorithm is used operationally in the PCASS instrument due to differences in instrumentation, underlying base rates, environmental factors, and cultural differences in the interviewees, among others.

Background

The Department of Defense has an urgent need for an improved ability to ascertain the truthfulness of individuals in the field during the routine conduct of its mission. To this end the Preliminary Credibility Assessment System (PCASS) was developed to collect physiological data during a structured interview process and to algorithmically evaluate the physiological data for changes inconsistent with truthfulness. The goal of the effort was not infallibility but rather a substantive improvement over pure intuition or other ad hoc techniques. In addition, the solution had to be highly portable, field operable, and require limited operator training.

The result of the effort is the PCASS instrument shown in Figure 1. The PCASS consists of a hand-held computer (Personal Digital Assistant or PDA) and a sensor assembly. Sensor data are communicated to the PDA across the Universal Serial Bus. The sensor assembly consists of two sensors. The first sensor measures perspiration through changes in galvanic skin resistance [1,2]. The second sensor measures cardiovascular changes using a photo-plethysmograph [2]. The sensor placement and the corresponding signals they collect are shown in Figure 2. The PDA contains software to construct a structured interview, guide the structured interview while recording physiological data, and evaluate the physiological data for truthfulness.

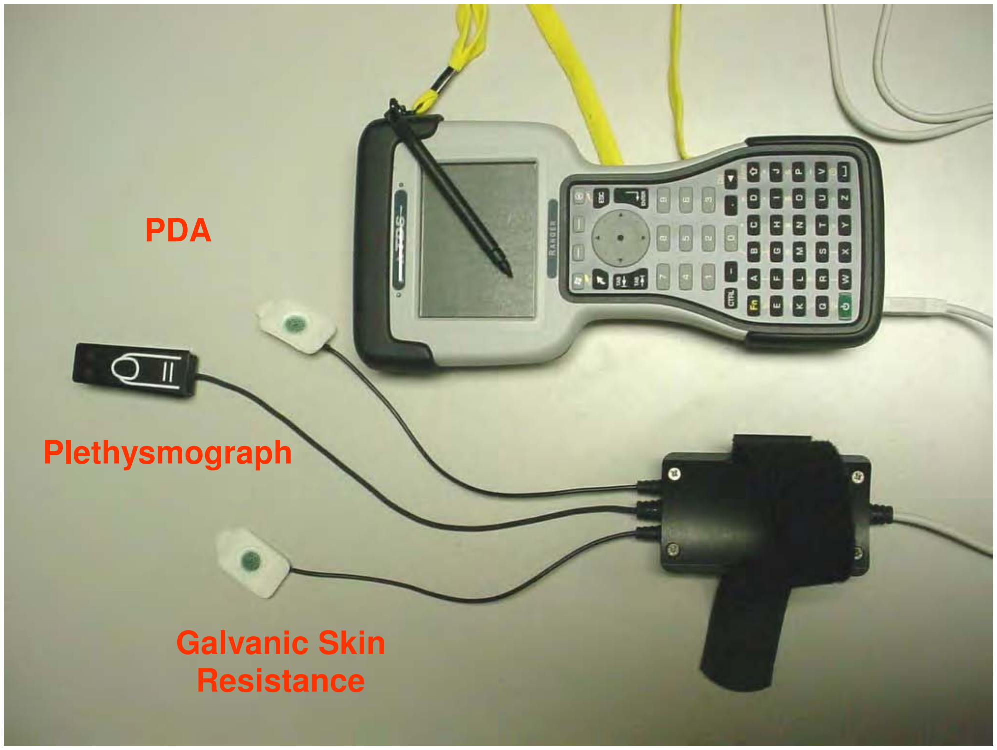


Figure 1. Preliminary Credibility Assessment System (PCASS)

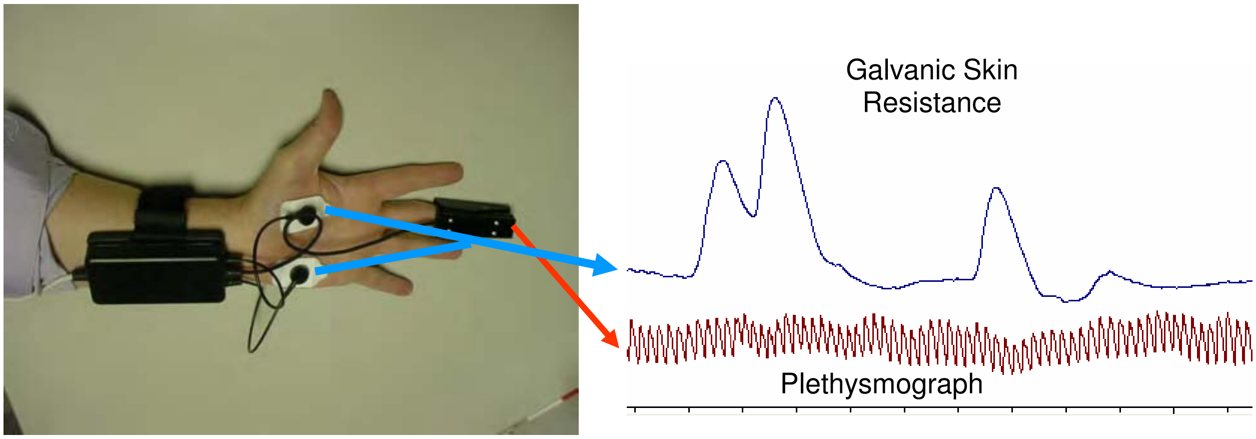


Figure 2. PCASS Sensor Assembly and Associated Signals

Specific Structured Interview Formats

The PCASS sensor data are collected in conjunction with the stimuli of questions presented during a specific structured interview. The format of this interview is similar to two of those used in specific issue polygraph testing, known as the Zone Comparison and Modified General Question Technique (MGQT) [4]. These formats alternate questions about the relevant specific issue with other control and irrelevant questions. One difference of the PCASS format is the ability to ask the questions in a continuous sequence, rather than breaking the interview into several “charts” separated by rest breaks. This is because traditional polygraph instruments collect cardiovascular data using an occlusive cuff which restricts blood flow and this can cause discomfort. This discomfort is relieved by deflating the cuff after each full presentation of all of the questions.

The PCASS allows for two interview formats according to the number of uniquely phrased relevant questions: a two relevant question format and a three relevant question format. The formats can be arranged two-dimensionally into four “rounds” of questions as shown in Tables 1 and 2 below. All of the presentations of a distinct relevant question (R1, R2, R3) constitute a “spot” and each spot is highlighted vertically below. Each format also includes three distinct control questions (C1, C2, C3), two irrelevant questions (I1, I2), and a sacrifice relevant (not evaluated) question (SR). Each relevant question is asked four times and each control question is asked at least three times. Only reactions to the control and relevant questions are processed by the evaluation algorithm. Sample questions are shown in Figure 3 and others can be found in Matte [4].

Table 1 — Structured Interview Format for Two Relevant Spots

I1	SR	C1	R1	C2	R2	I2
		C3	R1	C1	R2	I1
		C2	R1	C3	R2	I2
		C1	R1	C2	R2	C3

Table 2 — Structured Interview Format for Three Relevant Spots

I1	SR	C1	R1	C2	R2	C3	R3	I2
		C2	R1	C3	R2	C1	R3	I1
		C3	R1	C1	R2	C2	R3	I2
		C1	R1	C2	R2	C3	R3	

IRRELEVANT QUESTIONS
Are you sometimes called __________?
Is today __________?

SACRIFICE RELEVANT QUESTION
Regarding __________ do you intend to answer truthfully each question about that?

CONTROL QUESTIONS
[Before 200_,] Did you ever lie to someone who trusted you?
[Before 200_,] Did you ever lie to avoid responsibility for your actions?
[Before 200_,] Did you ever brag to impress others?

RELEVANT QUESTIONS
Did you plan with anyone to conduct the _______ attack on the United States?
Did you participate in any way in the _______ attack on the United States?
Did you ever meet with any of the people involved in the _______ attack on the United States?

Figure 3. Sample Questions

Methods

The PCASS instrument was in development concurrently with that of the embedded algorithm. Consequently, no PCASS data were available for development of the algorithm; though such data could become available at a later date for additional validation or development. Therefore, field polygraph data were used for algorithm training, testing, and validation. The polygraph data consist of 1392 examinations conducted by Department of Defense Polygraph Institute trained examiners from 1991 through 1997 and the data represent a variety of the commonly used *specific-issue* examination techniques: Zone Comparison, Modified General Question Technique (MGQT), Air Force MGQT, and BiZone. The data were collected using Axciton digital polygraph instruments with the collection of the electrodermal signal set to *automatic mode* (which uses hardware filtering). 784 of these exams were confirmed as either truthful (152) or deceptive (632). The remaining 608 exams were labeled by a consensus of the administering examiner and two, so-called, blind evaluators.

We divided the data 60% into a training set for algorithm development, 20% into a test set, and 20% into a final validation set according to the following scheme. To assure a more balanced distribution of easy and difficult cases to all sets, the data were first ordered by whether they were truthful or deceptive and then according to the probability of deception as determined by a commercially available evaluation algorithm (PolyScore® Zone/MGQT 5.5). Every second and fourth cases were assigned to the test and validation sets, respectively, and the others were assigned to the training set. The resulting training set consisted of 556 deceptive exams and 280 truthful exams. Of these, 374 were confirmed as deceptive and 88 were confirmed as truthful. Only confirmed data were retained for the test and validation sets. The training dataset was used to develop the algorithm. The test dataset was used occasionally during algorithm development to evaluate algorithm design alternatives. The validation set was evaluated only after the algorithm was finalized. When appropriate, a single, combined validation set was formed from the test and validation sets.

The sensor data in the polygraph database is similar but not identical to that which will be collected by the PCASS instrument. The electrodermal, or Galvanic Skin Resistance, signal collected by the Axciton is both hardware filtered and of lower resolution (12-bit vs. 18-bit) than that collected by the PCASS instrument. The filtering difference is accommodated by an equivalent digital filter. Note that this digital filter was originally designed for an earlier polygraph instrument which collected its electrodermal using conductance rather than resistance units. Though it does not affect our algorithm development, since we are using Axciton data directly, the PCASS signal should be reciprocated into conductance units before filtering. However, because of local linearity, after signal standardization there is frequently no discernable difference between the GSR resistance signal and GSC conductance signal, as shown in Figure 4. No processing is necessary to accommodate the increased resolution of the PCASS signal.

The cardiovascular data collected by a polygraph instrument such as the Axciton, uses a very different sensor than the PCASS instrument. The polygraph uses an occlusive cuff which uses pressure to measure changes in blood volume, typically of the upper arm. The PCASS instrument uses a photo-plethysmograph, which uses infrared light absorption to measure changes in blood volume at the finger-tip.

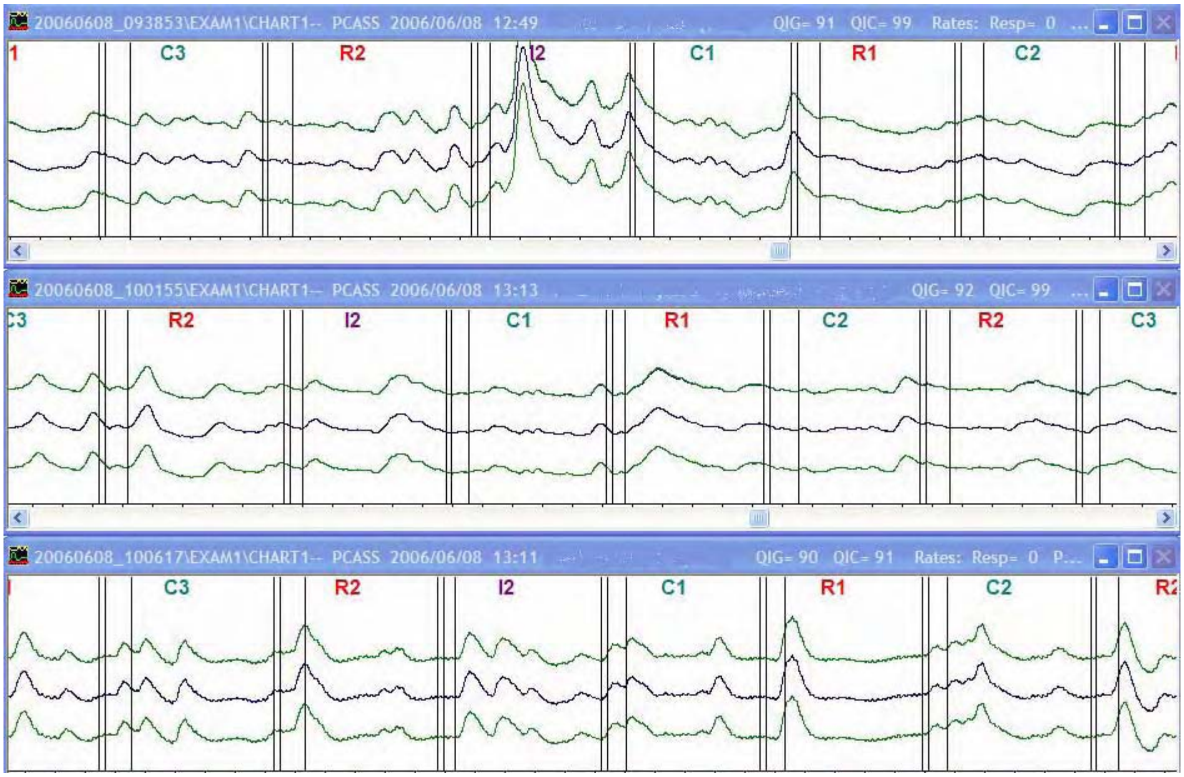

20060608_093853\EXAM1\CHART1-- PCASS 2006/06/08 12:49
QIG= 91 QIC= 99 Rates: Resp= 0 ...
1
C3
R2
I2
C1
R1
C2
20060608_100155\EXAM1\CHART1-- PCASS 2006/06/08 13:13
QIG= 92 QIC= 99 ...
C3
R2
I2
C1
R1
C2
R2
C3
20060608_100617\EXAM1\CHART1-- PCASS 2006/06/08 13:11
QIG= 90 QIC= 91 Rates: Resp= 0 P...
I
C3
R2
I2
C1
R1
C2
R2

Figure 4. GSR (Blue), GSC (Green) and Overlaid Signals (Top)

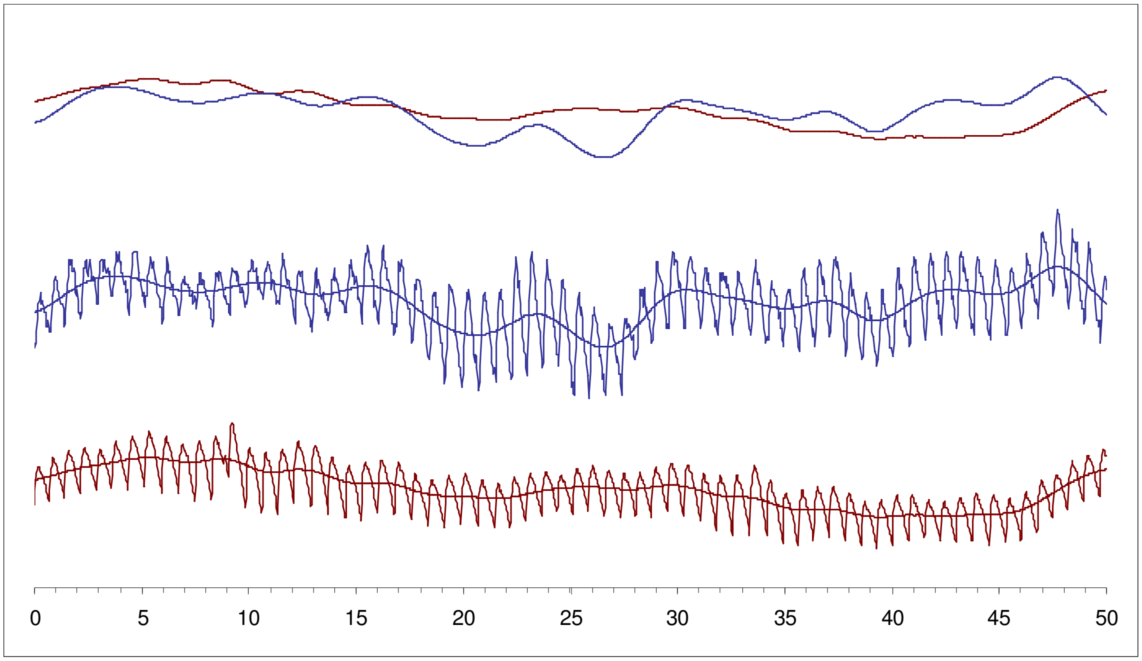


Figure 5. Cuff (Red) and Plethysmograph (Blue) Signals Overlaid With Bloodvolume and Both Bloodvolume Signals Also Overlaid (Top)

We had only a limited dataset containing both occlusive cuff and plethysmograph signals measured simultaneously from the same subject. As seen from the red cuff and blue plethysmograph signals in Figure 5, though the two signals are frequently similar, they are not identical. Thus, we did not have confidence that an algorithm developed on volume features would extrapolate from one sensor to the other. Fortunately, the pulsatile changes occurring at each heartbeat that are superimposed on these volume changes were in very good agreement in their heartbeat to heartbeat spacing, though not pulsatile amplitudes, as shown in Figure 6. Therefore we decided to use changes in inter-beat intervals to capture cardiovascular information. These inter-beat intervals were used to derive a so-called cardio tach signal [2]. More details of the cardio tach signal are provided in the Appendix.

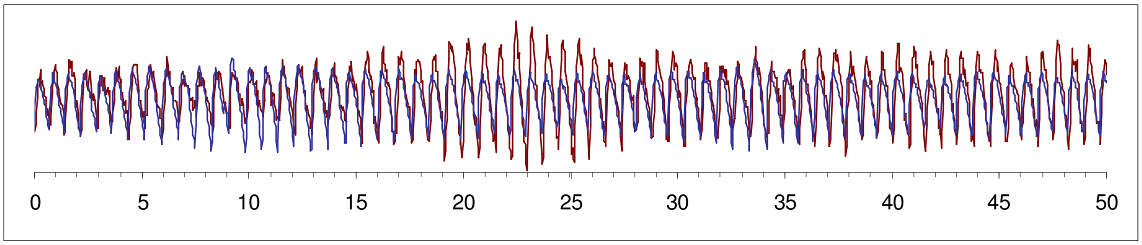
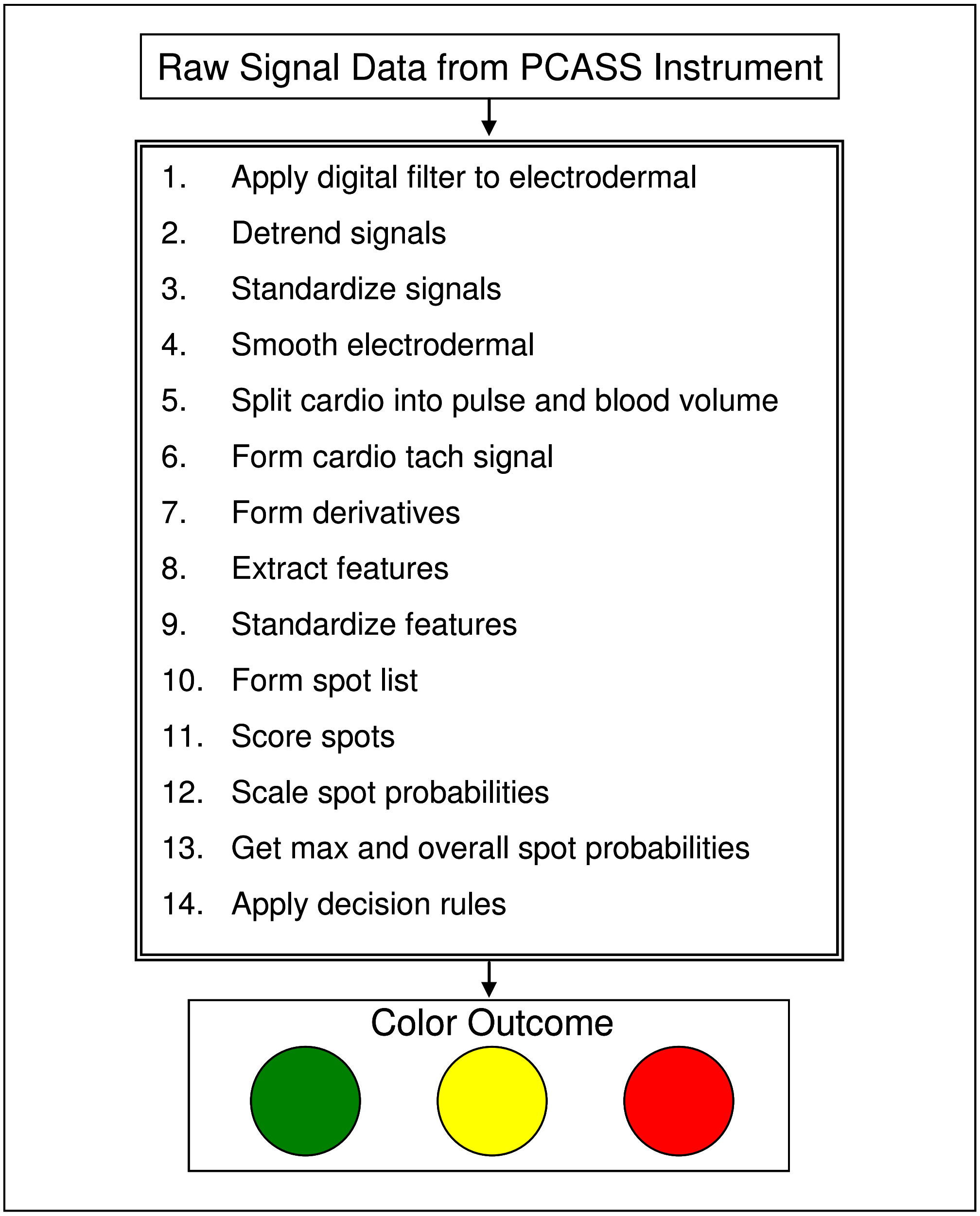


Figure 6. Cuff (Red) and Plethysmograph (Blue) Signals Overlaid Showing Good Agreement of Inter-beat Intervals

Algorithm and Processing Overview

At the completion of the interview, the raw signal data collected from the PCASS sensors are passed to the embedded algorithm for evaluation. From this raw signal data, a color-coded outcome is determined—red, yellow, or green. Figure 7 below provides a high-level description of the processing steps that transform the raw signal data into a color-coded outcome. The main steps involve signal detrending and standardizing; creating new signals; feature extraction and standardization; spot creation and outcome determination. Detrending of the raw electrodermal and cardiovascular signals removes the long (greater than 30-second) overall trends that are not question specific. Since the raw signal data from a population of individuals will vary widely in both scale and location, the detrended signals are standardized so that all electrodermal and cardiovascular signals have similar scales. The electrodermal and cardiovascular signals are then processed and split into different component signals. One important processed signal is the derivative signal which measures the rate of change in question reactions. Features related to those such as amplitudes and durations are extracted from each processed signal. Features from the relevant questions are standardized against the control questions by using the mean of the control questions and their pooled standard deviation. This step critically captures the differential changes occurring at the relevant questions compared to the control questions. It is from these standardized features that spot features are formed by simple averaging. The standardized relevant spot features are used as inputs to a Naïve Bayes classifier to produce spot probabilities. From these spot probabilities a decision rule is applied to determine the color coded outcome. All processing steps are discussed in further detail in the Appendix.



Raw Signal Data from PCASS Instrument

1. Apply digital filter to electrodermal
2. Detrend signals
3. Standardize signals
4. Smooth electrodermal
5. Split cardio into pulse and blood volume
6. Form cardio tach signal
7. Form derivatives
8. Extract features
9. Standardize features
10. Form spot list
11. Score spots
12. Scale spot probabilities
13. Get max and overall spot probabilities
14. Apply decision rules

Color Outcome

Figure 7. PCASS Processing

Spot Evaluation

A spot is all the repetitions of a particular phrasing of a relevant question and spot evaluation is fundamental to the algorithm. The features for each relevant question repetition are averaged into relevant spot features after first being standardized against statistics from all of the control questions. Depending on the particular questions, an interviewee could be lying to one relevant question while telling the truth to the others. To allow for this possibility, each distinct spot is evaluated separately and the spot with the largest reaction (maximum spot probability) on the interview is identified and used in determining the outcome. In addition, to gain power from the case where the interviewee is lying to more than one question, all of the relevant questions on each interview are combined into a single, overall "spot" which is also evaluated. Only actual spots were used to train the classifier.

Naïve Bayes Spot Classifier

A limited experimentation with a variety of classifier types including logistic regression, support vector machines, classification trees, and neural networks, was done before the Naïve Bayes classifier was selected due to its simple form and robustness. The fundamental idea behind the Naïve Bayes classifier is to compare how well an interview matches the estimated distributions of the truthful and deceptive populations and to produce a probability that an interviewee is deceptive. The Naïve Bayes classifier is related to traditional quadratic discriminate analysis with the exception that the covariance matrices are constrained to be diagonal. This reduces the form of the classifier to a simple product of density functions [11,12].

For our Naïve Bayes spot classifier development, all of the relevant spots from the 836 cases in our training set were used. The 556 deceptive interviews resulted in 1777 deceptive spots. The 280 truthful interviews resulted in 804 truthful spots for a total of 2581 spots. All of the training set spots were used for both feature selection and parameter estimation.

The Naïve Bayes spot classifier development began with a total feature set of 441 features. This feature set is comprised of various percentiles, percentile differences, and time to percentiles averaged over each reaction for each spot (see Appendix). From this feature set, stepwise selection was performed and the model that maximized the binomial likelihood (as is also used in logistic regression [7]) was retained. We investigated two families of density functions: Normal and Cauchy. The heavy tails of Cauchy distribution limits the effect of unusual observations; in particular for a single unlikely feature which would tend to drive the product very close to zero. Compared to using the more commonly used Normal distribution, we found that the Cauchy distribution stabilized the overall probabilities from a computational standpoint and resulted in a classifier which extrapolated better from the training set to the test set. All that is needed to train a Naïve Bayes classifier are the sample statistics computed for each feature for the deceptive and truthful groups.

The 26 features identified by the stepwise procedure are shown in Table 3. The Signal column shows from which sensor or derived signal the feature is calculated. The processing column shows additional processing done to the signal. The Start and Stop columns show the period of data used after the beginning of each control and relevant question, in seconds. The Question Feature column shows the information extracted from the signal during the given period. The Mean and Standard Deviation columns show the sample statistics.

Table 3 — PCASS Spot Model

Signal	Processing	Start	Stop	Question Feature	Truthful		Deceptive	
					Mean	Std. Dev.	Mean	Std. Dev.
Cardio Tach	derivative	1.5	9.5	85th percentile	0.0565	0.8748	-0.0437	0.8334
Cardio Tach	derivative	1.5	9.5	90th percentile	0.0414	0.8789	-0.0329	0.8404
Cardio Tach	derivative	1.5	9.5	95th percentile	0.0165	0.8721	-0.0243	0.8373
Cardio Tach	derivative	1.5	9.5	Time to 45th percentile	-0.0759	0.8781	0.2094	0.8026
Cardio Tach	derivative	1.5	9.5	Time to maximum	-0.0046	0.9374	-0.1656	0.8434
Cardio Tach	derivative	1.5	9.5	Maximum	0.0024	0.8628	-0.0214	0.8477
Cardio Tach	derivative	1.5	9.5	55th - 45th percentiles	-0.0474	0.8078	0.1442	0.7990
Cardio Tach	derivative	1.5	9.5	90th - 85th percentiles	-0.0362	0.8086	0.0188	0.8227
Cardio Tach	derivative	1.5	9.5	Time between 50th and 25th percentiles	-0.0680	0.8325	-0.0526	0.8276
Cardio Tach		1.5	9.5	65th percentile	-0.1449	0.9035	-0.1611	0.9345
Cardio Tach		1.5	9.5	70th percentile	-0.1517	0.9015	-0.1455	0.9030
Cardio Tach		1.5	9.5	75th percentile	-0.1558	0.8992	-0.1313	0.8953
Cardio Tach		1.5	9.5	80th - 75th percentiles	-0.1481	0.8815	-0.1231	0.8947
Cardio Tach		1.5	9.5	80th percentile	-0.1462	0.8668	-0.1182	0.8935
Cardio Tach		1.5	9.5	Time to 50th percentile	-0.0817	0.7866	-0.0322	0.8047
Cardio Tach		1.5	9.5	Time between 95th and 5th percentiles	0.1226	1.0363	-0.3520	0.9242
Cardio Tach		1.5	9.5	Minimum	-0.0045	0.8735	-0.3210	0.8645
Cardio Tach		1.5	9.5	85th - 75th percentiles	-0.0690	0.7729	0.0370	0.8950
Cardio Tach		1.5	9.5	85th percentile	-0.0524	0.7956	0.0089	0.9145
Electrodermal	Automatic mode	1	13	70th percentile	-0.5360	0.9873	1.1610	1.0553
Electrodermal	Manual mode	1.5	14	65th - 15th percentiles	-0.5661	1.0405	1.1029	1.0820
Electrodermal	Manual mode	1.5	14	Time between 75th and 50th percentiles	-0.3989	0.9763	0.1413	0.9592
Electrodermal	Manual mode derivative	3	10	Time to 35th percentile	0.1888	1.2504	0.2024	1.0156
Electrodermal	Manual mode derivative	3	10	Time to 50th percentile	0.0629	0.9633	0.0745	0.9478
Electrodermal	Manual mode derivative	3	10	Time between 75th and 50th percentiles	-0.2763	0.9524	-0.1976	0.8902
Electrodermal	Smoothed automatic mode	1	13	Time between 75th and 50th percentiles	-0.1593	0.9183	0.1191	0.7667

Color-Coded Outcome Decision Rules

Once the Naïve Bayes Classifier was finalized, decision rules were derived to produce color-coded outcomes. Determining these decision rules is both non-trivial and subjective. They depend on both the maximum and overall spot probabilities and produce one of three possible outcomes—red, yellow, or green. These respectively indicate changes associated with deceptiveness, truthfulness, or otherwise, inconclusiveness. There is a tradeoff between more inconclusive outcomes and overall accuracy. There is also a tradeoff between false positive and false negative outcomes. The rules consist of thresholds for both the maximum and overall spot probabilities above which an interview is classified as red and below which it is classified as green. The default is yellow. We explored a wide range of possible decision rules and in consultation with the Department of Defense Counterintelligence Field Activity (CIFA) we selected the set which agreed best with their desire to bound the overall inconclusive rate at 30% and to minimize false negative errors. The resulting rules were developed entirely from our training set and are shown graphically in Figure 12 below.

Results

In this section we provide several performance results. We provide Receiver Operating Characteristic (ROC) results for the Naïve Bayes Classifier as applied to both the maximum and overall spots. We provide outcome decision accuracy for the embedded algorithm as a whole, which makes use of the Naïve Bayes probabilities and the outcome decision rules. We also examine the sensitivity of the results to the underlying, and almost certainly unknown, base rate of deception.

Maximum and Overall Spot ROC Results

Outputs from the Naïve Bayes classifier for maximum and overall spots were used to generate Receiver Operating Characteristic (ROC) curves [5,6] for the training, test, and validations datasets. Also, ROC curves are shown for the combined test and validation data. The motivation and justification for combining these two datasets are provided in a subsequent section. ROC curves can be used to assess classifier performance without the need for specific classifications. The area under the ROC curve (AUC) is equivalent to the pair-wise concordance of the classifier outputs (and also to the Wilcoxon/Mann-Whitney statistic). That is, AUC is the probability that given both a truthful and a deceptive interviewee, the deceptive interviewee has the larger value for his spot. Figure 8 shows the ROC curves produced by the training data maximum (AUC=0.916) and overall spots (AUC=0.927). Figure 11 shows the ROC curve for the combined validation data maximum (AUC=0.934) and overall spots (AUC=0.945). The 95% confidence intervals for the areas are shown in figures in parentheses. From these it can be seen that though there was a slight improvement of the AUC values on the validation data over the training data, the training and validation performances are not statistically significantly different.

The maximum and overall spots are obviously correlated, as the overall spot includes data used in the maximum spot. The Pearson correlation coefficient for the confirmed training set is 0.86 and for the combined validation set the Pearson correlation coefficient is 0.84. Figure 12 shows a scatter plot of the maximum and overall spots for the training data. From this it can be seen that it is unusual for the overall spot probability to exceed the maximum spot probability, as might be expected. It can be seen from the lower right region of the figure, in particular, the benefit derived from the use of both spot probabilities.

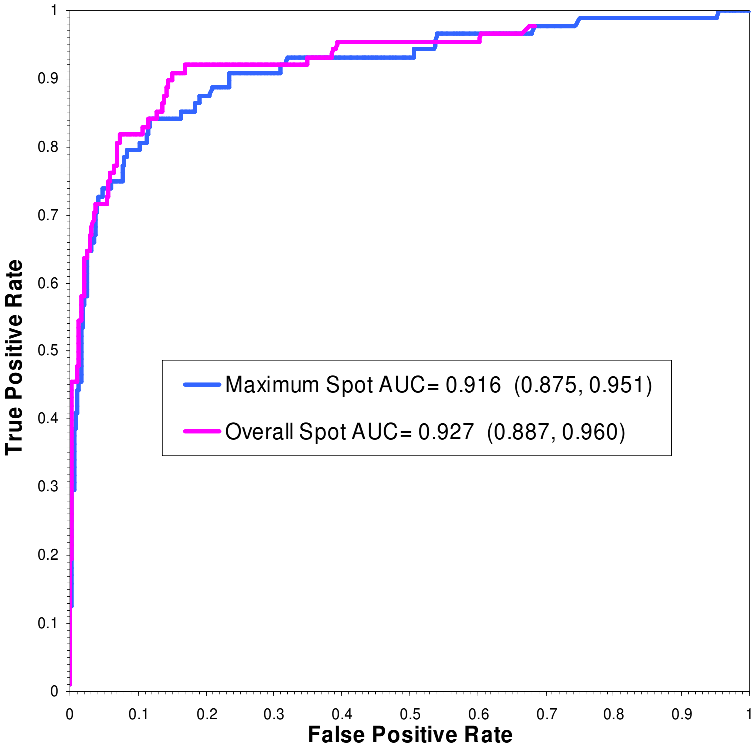


Figure 8. ROC Curves of Confirmed Training Data with Area Under the Curve (AUC) and 95% Confidence Bounds

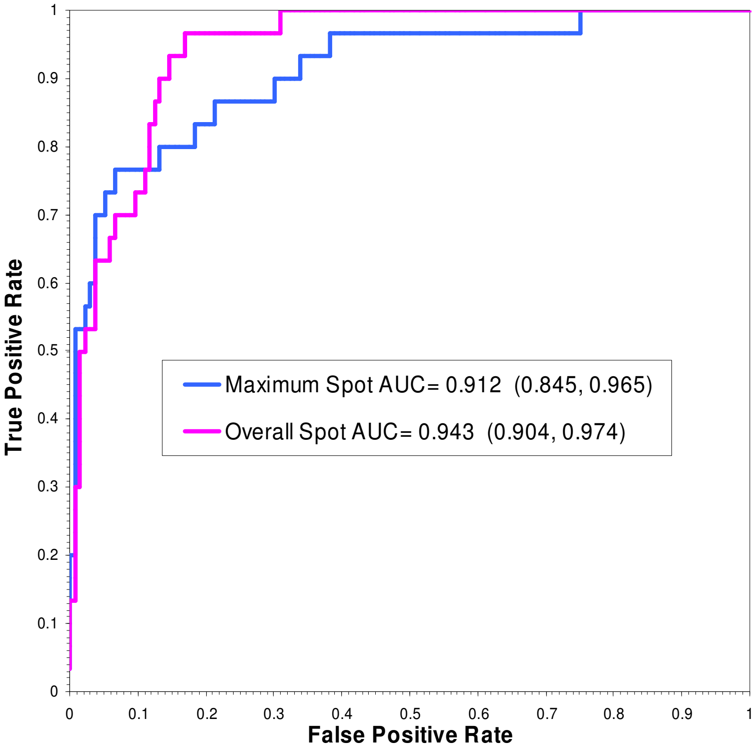


Figure 9. ROC Curves of Test Data with Area Under the Curve (AUC) and 95% Confidence Bounds

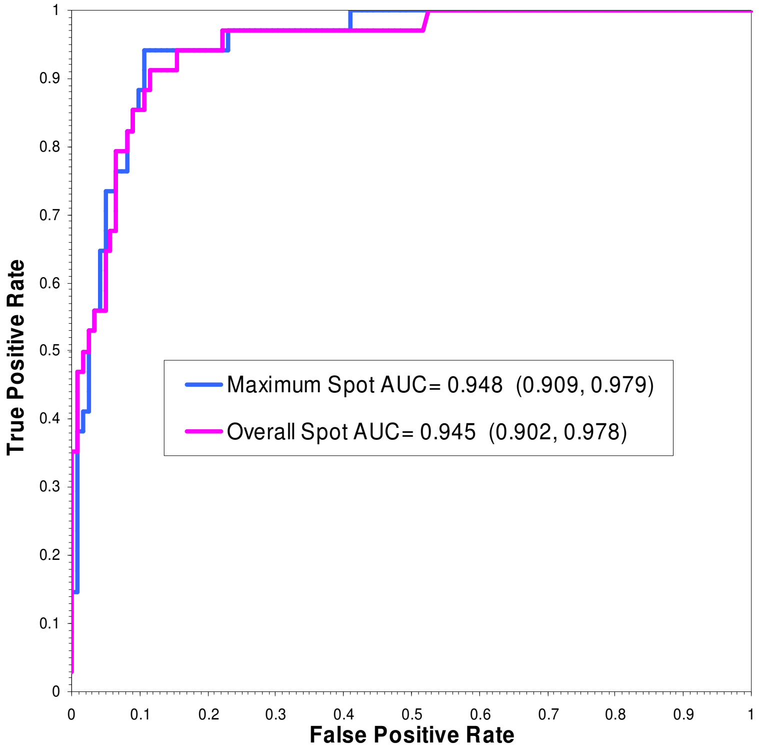


Figure 10. ROC Curves of Validation Data with Area Under the Curve (AUC) and 95% Confidence Bounds

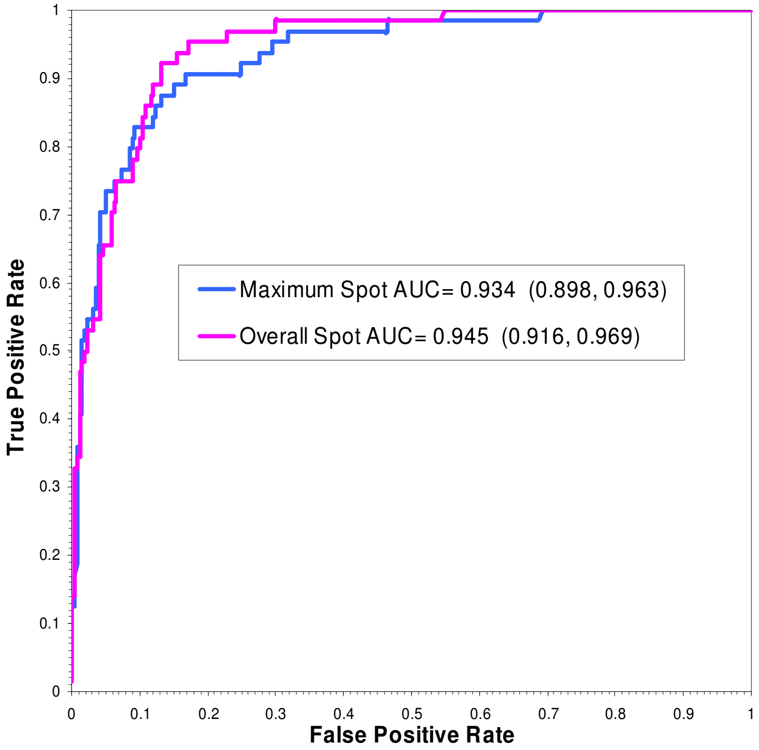


Figure 11. ROC Curves of Combined Validation Data with Area Under the Curve (AUC) and 95% Confidence Bounds

Outcome Decision Rules

When evaluating an interview, the Naïve Bayes probabilities for both maximum and overall spots are used to classify the interview outcome as red, yellow, or green according to where the spot values jointly fall. The decision rules which assign the outcomes were developed on the training data maximum and overall spot probabilities with the intent of keeping the number of yellow outcomes below 30% and also of minimizing false negative outcomes. The decision rules are shown graphically as colored regions in Figure 12. In this figure, the blue diamonds correspond to truthful interviewees and the red squares correspond to deceptive interviewees. Yellow outcomes have the effect of increasing the proportion of red outcomes that are deceptive and increasing the proportion of green outcomes that are truthful, at a cost in utility. Operationally a yellow outcome is treated as inconclusive and indicates that further interviewing is needed.

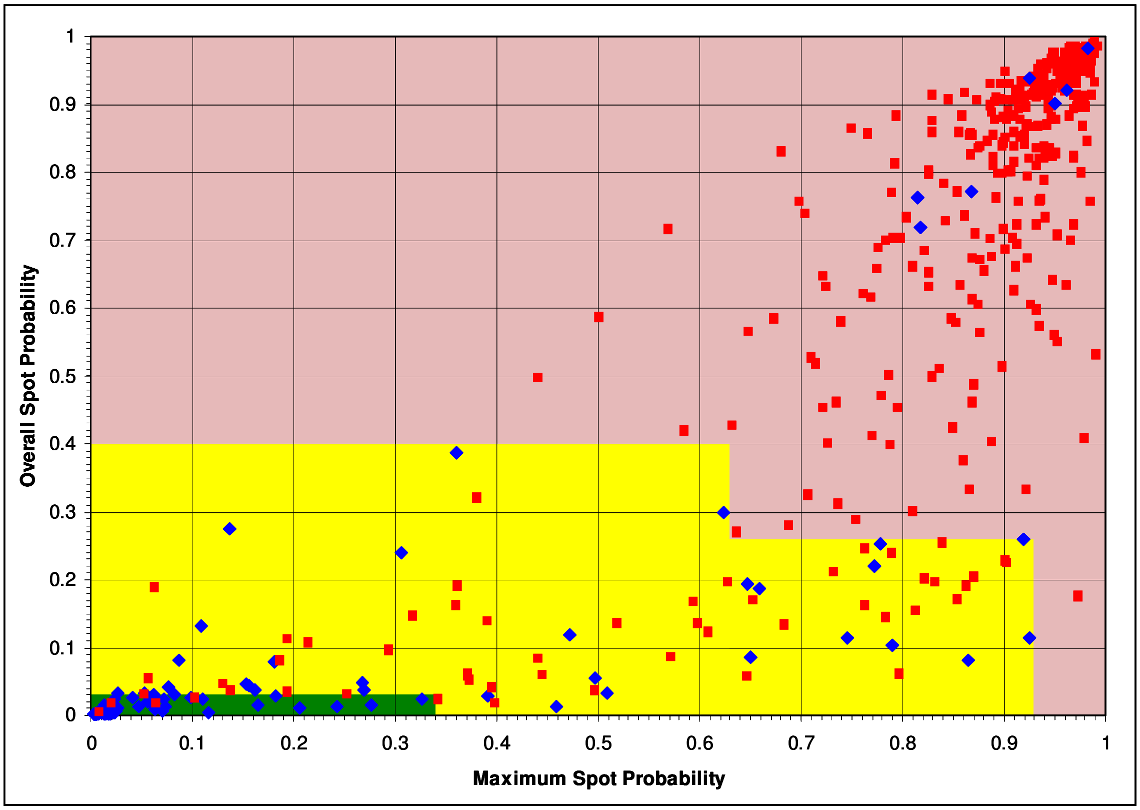


Figure 12. Outcome Decision Rules Overlaid on Confirmed Training Data

Outcome Results

The results of applying the algorithm to the training, test and validation datasets are shown for the confirmed truthful interviews in Table 4 and the confirmed deceptive interviews in Table 5. Both actual counts and percentages are shown. The parenthesized values represent the 95% confidence interval. Based on Fisher's Exact test, there is no statistically significant association between the training, test and validation datasets and outcomes for either the confirmed truthful (p= 0.9001) or confirmed deceptive (p=0.9249) groups. This demonstrates the algorithm's ability to generalize to new data similar to that used for training. Furthermore, based on the limited use of the test dataset, it would be expected that there would be no difference in performances on the test and validation datasets. Once again, based on Fisher's Exact test, there is no statistically significant difference between the test and validation datasets and outcomes for either the confirmed truthful (p= 0.8048) or confirmed deceptive (p=0.7536) groups. Therefore, because of the relatively small sample sizes of the confirmed truthful sets, we elected to combine the test and validation datasets into a single combined validation set for reporting our additional validation results. These combined results are shown graphically in the next section.

Table 4 — Confirmed Truthful Outcomes For Each Dataset

Dataset	Red	Yellow	Green	%Red	%Yellow	%Green
Training	7	32	49	8 (3, 17)	36 (24, 49)	56 (42, 68)
Test	2	14	14	7	47	47
Validation	3	13	18	9	38	53
Combined	5	27	32	8 (2, 20)	42 (28, 57)	50 (35, 65)

Table 5 — Confirmed Deceptive Outcomes For Each Dataset

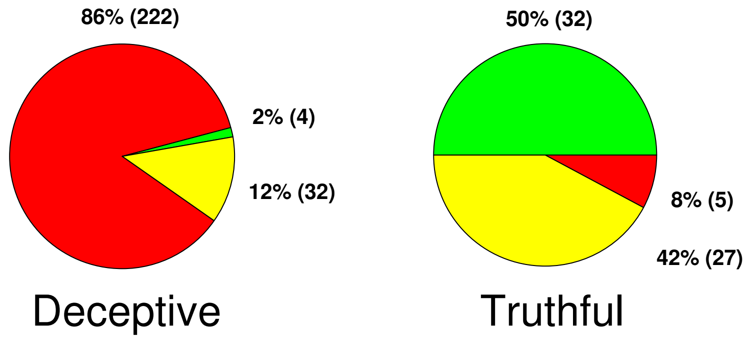
Dataset	Red	Yellow	Green	%Red	%Yellow	%Green
Training	321	48	5	86 (81, 91)	13 (8, 17)	1 (0, 4)
Test	115	19	2	85	14	1
Validation	107	13	2	88	11	2
Combined	222	32	4	86 (80, 91)	12 (8, 18)	2 (0, 4)

Combined Validation Accuracy Results

The proportions of red, yellow, and green outcomes for both truthful and deceptive interviewees are shown in Figure 13 for the combined confirmed field validation data. We can determine accuracy for deceptive interviewees by defining red to be a correct outcome, green an incorrect outcome, and yellow an inconclusive outcome. The deceptive false negative error rate of 2% has a 95% confidence interval from nearly zero to 4%. Likewise, we can determine accuracy for truthful interviewees by defining green to be a correct outcome, red an incorrect outcome, and yellow an inconclusive outcome. The truthful false positive error rate of 8% has a 95% confidence interval from 2% to 20%. While it is possible to choose decision rules which balance false positive and negative error rates, the objective of minimizing false negative errors while controlling the inconclusive rate results in a higher percentage of false positive errors.

Combined Validation Results

Confirmed Field Cases



- **For 100 Deceptive Interviewees**
  - **86 Result in Red**
  - **12 Result in Yellow**
  - **2 Result in Green (False Negative)**
- **For 100 Truthful Interviewees**
  - **50 Result in Green**
  - **42 Result in Yellow**
  - **8 Result in Red**

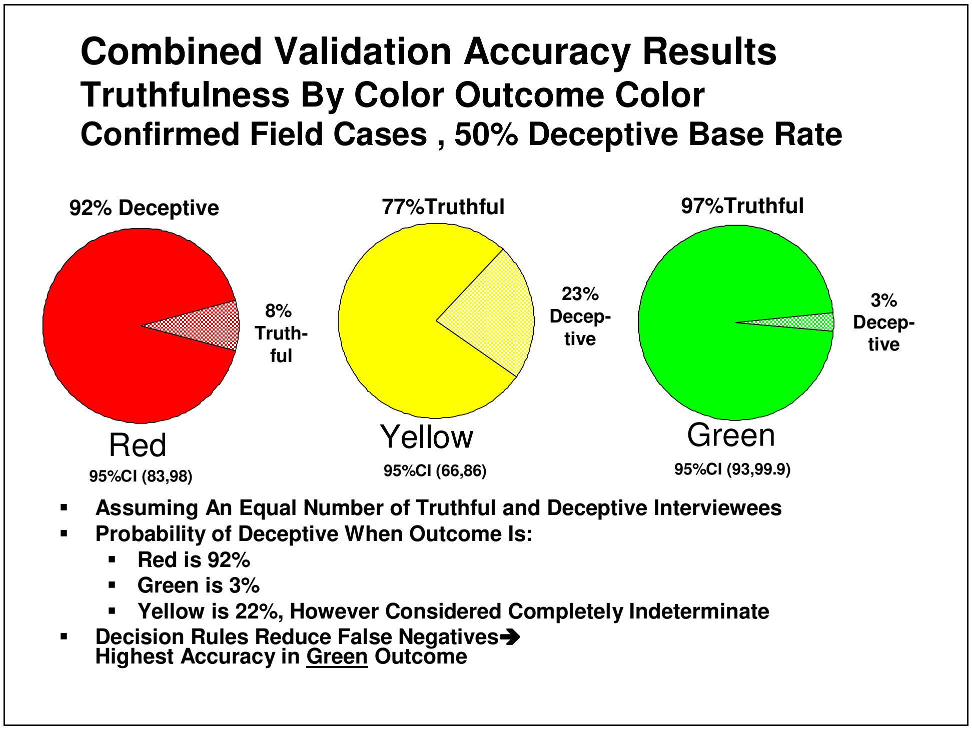
Figure 13. Color Outcome By Truthfulness For Combined Validation

Sensitivity to Base Rate of Deception

Though it is possible to describe performance results when it is known whether the interviewee is truthful or deceptive, as are shown in Figure 13, in the field it is precisely the purpose of the interview to determine whether or not the interviewee is truthful. The accuracy of an inference from outcome color to truthfulness is less well defined since it depends on the underlying relative proportion of truthful and deceptive interviewees—the so-called base rates and these are almost certain to be unknown. In this way, the results in Figure 13 are actually the degenerate cases for interviewing either a population that is entirely deceptive, in which case all green outcomes represent false negatives, or testing a population that is entirely truthful, in which case all red outcomes represent false positives. In actual field use there will be a mixture of red outcomes which consist of both true and false positives and green outcomes which consist of both true and false negatives. These mixtures depend both on the accuracies shown in Figure 13 and the deceptive base rate.

Figures 14-16 show outcome results for three base rates of deception: 50%, 90% and 10%; which cover the cases of a population which is half deceptive and half truthful, a mostly deceptive population, and a mostly truthful population. From these figures it can be seen that the percentage of red outcomes that are actually deceptive can vary from 99% for the mostly deceptive 90-10 base rate, to 92% for a 50-50 base rate, and to 55% for a mostly truthful 10-90 base rate. This trend shows what might be expected, that as the deceptive base rate decreases, the proportion of false positives increases. Conversely, the percentage of green outcomes that are truthful varies from 99.7% for a mostly truthful 10-90 base rate, to 97% for a 50-50 base rate, and to 78% for a mostly deceptive 90-10 base rate. This trend also shows what might be expected, that as the deceptive base rate increases, the proportion of false negatives also increases. It is important to point out that in the case of a mostly truthful population, a green outcome is rarely not a truthful interviewee. Likewise, in the case of a mostly deceptive population, a red outcome is very infrequently not a deceptive interviewee.

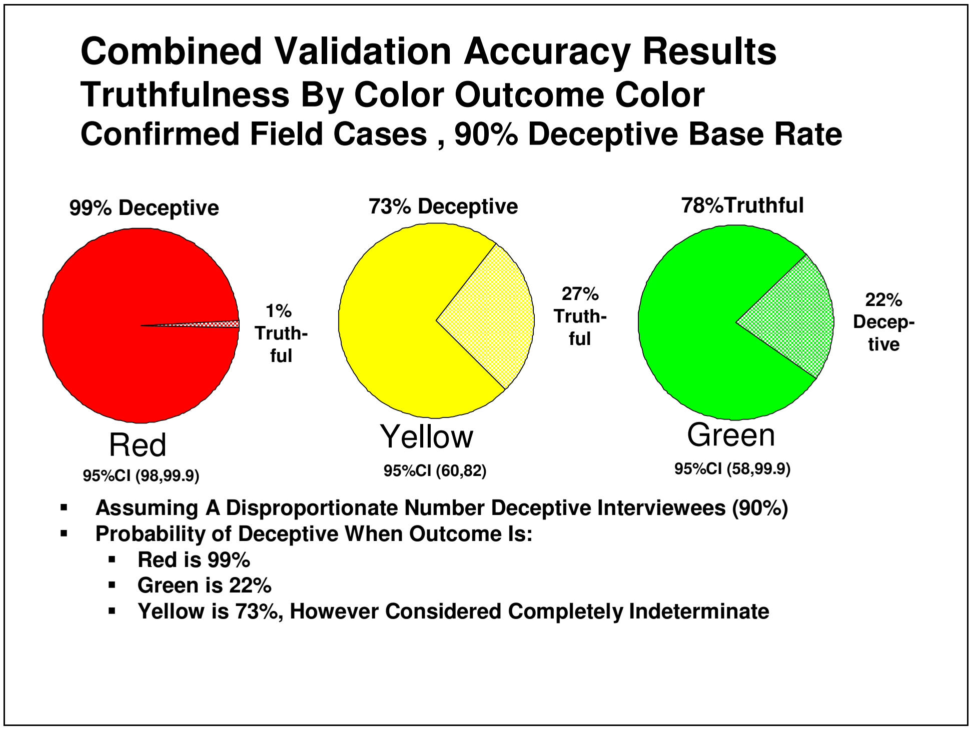
The effect of minimizing false negatives can be seen in the 50-50 base rate case, where 97% of the green outcomes are truthful, but a lesser percentage, 92%, of the red outcomes are deceptive. This effect is magnified by the larger proportion of truthful interviewees in the 10-90 base rate case, where even a yellow outcome has a strong likelihood, 97%, of being truthful—though this interpretation is strongly discouraged since the base rate of deception is very unlikely to be known. The effect of allowing yellow inconclusive outcomes can be seen in the 50-50 base rate case where, despite having 86% red outcomes on deceptive interviewees (Figure 13) and 50% green outcomes on truthful interviewees (also Figure 13), the accuracy of interpreting a red outcome as deceptive is 92% and a green outcome as truthful is 97%.



Combined Validation Accuracy Results
Truthfulness By Color Outcome Color
Confirmed Field Cases , 50% Deceptive Base Rate

- **Assuming An Equal Number of Truthful and Deceptive Interviewees**
- **Probability of Deceptive When Outcome Is:**
  - **Red is 92%**
  - **Green is 3%**
  - **Yellow is 22%, However Considered Completely Indeterminate**
- **Decision Rules Reduce False Negatives➔ Highest Accuracy in <u>Green</u> Outcome**

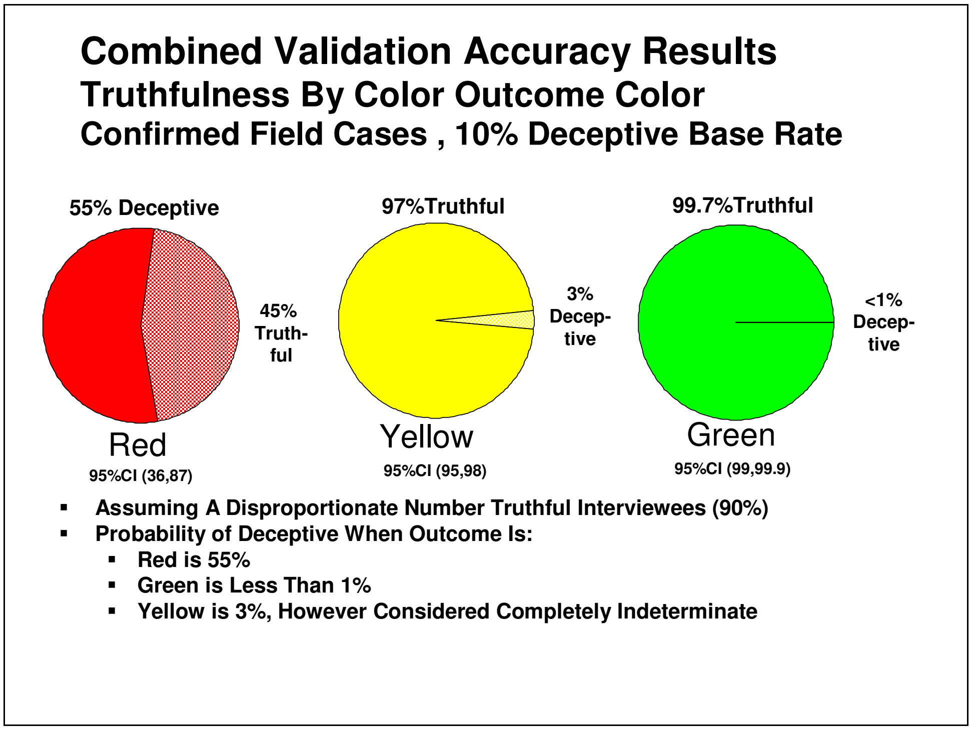
Figure 14 Truthfulness By Color Outcome For 50-50 Base Rate For Combined Validation



Combined Validation Accuracy Results
Truthfulness By Color Outcome Color
Confirmed Field Cases , 90% Deceptive Base Rate

- Assuming A Disproportionate Number Deceptive Interviewees (90%)
- Probability of Deceptive When Outcome Is:
  - Red is 99%
  - Green is 22%
  - Yellow is 73%, However Considered Completely Indeterminate

Figure 15. Truthfulness By Color Outcome For Mostly Deceptive 90% Deceptive-10% Truthful Base Rate For Combined Validation



Combined Validation Accuracy Results
Truthfulness By Color Outcome Color
Confirmed Field Cases , 10% Deceptive Base Rate

- Assuming A Disproportionate Number Truthful Interviewees (90%)
- Probability of Deceptive When Outcome Is:
  - Red is 55%
  - Green is Less Than 1%
  - Yellow is 3%, However Considered Completely Indeterminate

Figure 16. Truthfulness By Color Outcome For Mostly Truthful 10% Deceptive-90% Truthful Base Rate For Combined Validation

Sensitivity to Other Factors

In addition to sensitivity to base rate there are other factors which may influence field accuracy. These include:

- Unrecognized differences between the data collected by a polygraph instrument, such as we used for algorithm development and that collected by the PCASS.
- Differences between polygraph examiners and the actual PCASS operators due to both their backgrounds (e.g. law-enforcement versus military) and their training.
- Cultural differences in the population being interviewed with respect to the perceived consequences of lying. Such differences affect the selection and phrasing of questions, and the pre-interview instructions.
- Differences in the operating environment in terms of ambient temperature, humidity, and noise (e.g. an air-conditioned, sound-proofed room versus the desert).

Another accuracy issue is related to the use of confirmed field deceptive data for reporting results. This is because the confirmations are not entirely independent of the polygraph exams themselves and therefore may under represent false negative errors. These are all factors which we cannot evaluate with our validation data. Many of these can be addressed by future studies using the PCASS instrument in an operationally realistic setting. Some of these may already have been addressed by using polygraph under similar conditions.

Summary

In summary, we have produced an initial algorithm for the evaluation of the PCASS data and have demonstrated its ability to generalize to independent polygraph validation data. The algorithm was specifically designed to minimize false negative errors, that is, deceptive interviewees resulting in green outcomes, which results in a greater number of truthful interviewees with either yellow or red outcomes. In addition, the algorithm was designed to use features for which there is confidence that they will extrapolate to the PCASS instrument. In particular, features based on time rather than amplitudes were selected to extrapolate from the cuff to the photo-plethysmograph. It still remains to perform a validation of the algorithm on the PCASS instrument itself in an operationally realistic setting. Once sufficient PCASS units are produced and a relatively large database, say of roughly 250 truthful and 250 deceptive interviews, is acquired then an algorithm could be developed which further exploits the plethysmograph signal and may allow the use of volume features. This could result in some combination of improved accuracy and a reduced rate of inconclusive outcomes.

Appendix—Processing and Algorithm Overview

Over the course of the past two decades The Johns Hopkins University Applied Physics Laboratory (APL) has developed and fielded a highly successful algorithm for objectively evaluating digitally collected ZONE/MGQT polygraph data [8]. This algorithm employs a variety of novel, statistical techniques for characterizing the polygraph reactions being evaluated. These characterizations are referred to as features and they are fundamentally different from those that have been traditionally used for polygraph evaluation and from those that have been reported in the psychophysiology and polygraph literature [1,2,3]. They are primarily based on the concept of percentiles, which measure how large a reaction is for how long.

Our features are based on a subset of the same raw signal data as have been used in traditional digital polygraph for many years. These signals include changes in the electrical conductivity of the skin (electrodermal) due to perspiration and a cardiovascular measurement which is provided by a standard blood pressure cuff. In the polygraph setting, the cuff measures changes in overall blood volume and pulse rather than blood pressure. In the PCASS setting, the cuff is replaced by a photo-plethysmograph.

Polygraph Data

Features traditionally relate to changes in the signals, or reactions, at each question of interest—the control and relevant questions. These features often relate to either the time or amplitude axes; such as the duration of a reaction or its maximum amplitude. Other features include rate changes [3]. Figure A-15 shows some of the kinds of information that the features are intended to capture.

Modern polygraph equipment retains these same measurements, but collects the data digitally rather than in an analog fashion on paper charts. Figure A-1 depicts a set of digitized signals. For our purposes, these data are collected at 30 samples per second. In the case of the Axciton digital polygraph instrument, the data are collected using a 12-bit analog-to-digital (A/D) converter. This results in the data values falling in a range of 0-4095. The PCASS unit collects at a higher 18-bit resolution resulting in a data range of 0-262143.

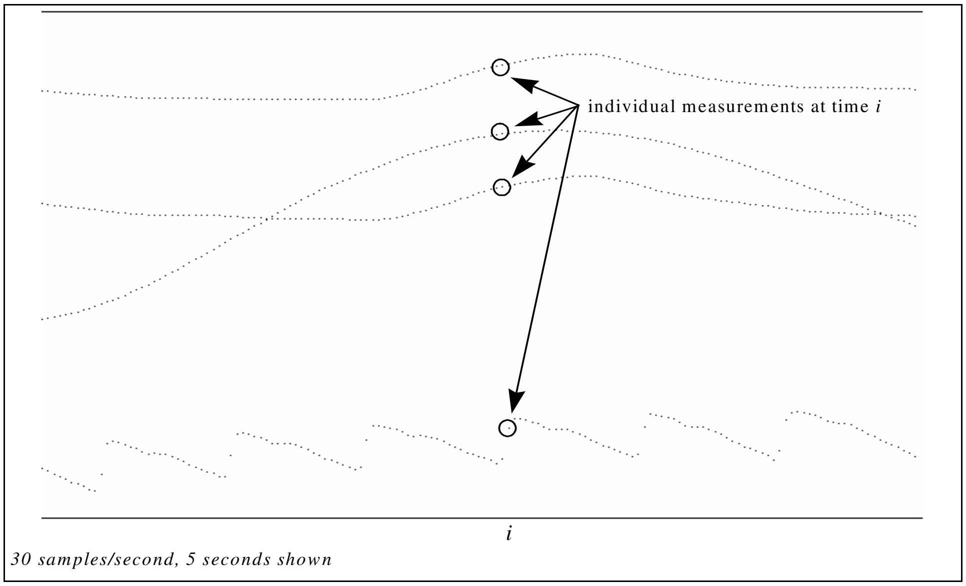


Figure A-1. Digitized Polygraph Signals

The mapping of signal amplitudes into this range is accomplished by the sensitivity settings of the instrument. Thus, one interviewee's values may move through a range of 100-200 during a particular interview while another interviewee's may span 1000-4000. The range of a signal's values is a function of both the interviewee's physiology and the collection instrument settings. Thus, the data lack consistent amplitude units; a problem that is addressed by signal standardization.

Algorithm Overview

Our algorithm ultimately employs a Naïve Bayes classifier to discriminate deceptive from truthful interviewees. The classifier produces the probability that a given set of features belongs to an interviewee who is deceptive. Much of our early work was concerned with the development of useful features from which to determine which of the two populations (deceptive and non-deceptive) the interviewee's measurements belong. Figure 7 lists the main processing steps and Figure A-2 shows the overall processing flow of the basic scoring algorithm. As can be seen from Figure A-2, much of the algorithm involves processing the raw digitized data to produce other derived signals. These derived signals make reactions easier to characterize with features. There are several kinds of processing involved with deriving these signals and each has a different purpose:

- *Detrending:* Remove long-term trends. Long-term trends are not related to a particular question response.
- *Cardio Splitting:* Isolate high frequencies corresponding to pulse from low frequencies corresponding to overall blood volume changes.
- *Cardio Tach:* Identify individual heart beats and changes in heart rate.
- *Automatic to Manual (or vice versa):* allow benefits of both filtered and unfiltered electrodermal.
- *Derivative:* Isolate how quickly a reaction is changing.

As a final part of the signal processing, the signals are standardized. Each of these steps is described in more detail below.

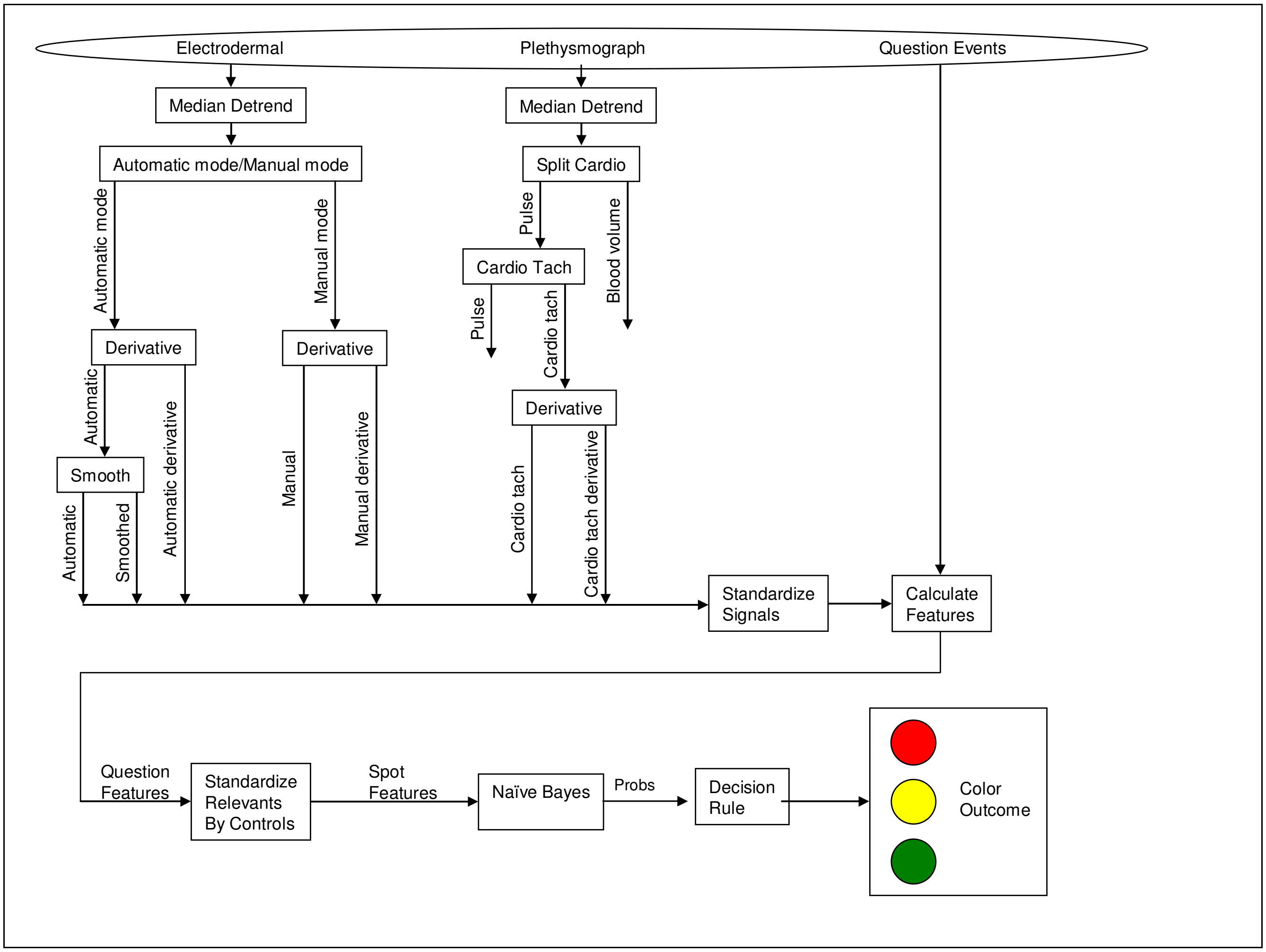


Figure A-2. Conceptual Algorithm Overview

Automatic-Manual Mode Electrodermal Conversions

The electrodermal signal as collected by the Axciton digital polygraph instrument is expected to be collected with the Axciton sensor box in the so-called automatic mode. This method of collection uses a hardware filter which attempts to keep the electrodermal signal at a certain nominal value. Excursions by the signal, either up or down, are drawn back to this nominal value. This filtering tends to sharpen the overall shapes of the reactions and maximizes the use of the 12-bit range of the A/D converter.

The PCASS instrument collects an unfiltered (manual mode) electrodermal signal using a 18-bit A/D converter. Our algorithm uses a digital software filter to transform the PCASS electrodermal to an Axciton automatic mode equivalent. Likewise, we can transform the Axciton to an approximate manual mode equivalent. Both signals are available for feature extraction. Figure A-3 depicts the same signal in both modes.

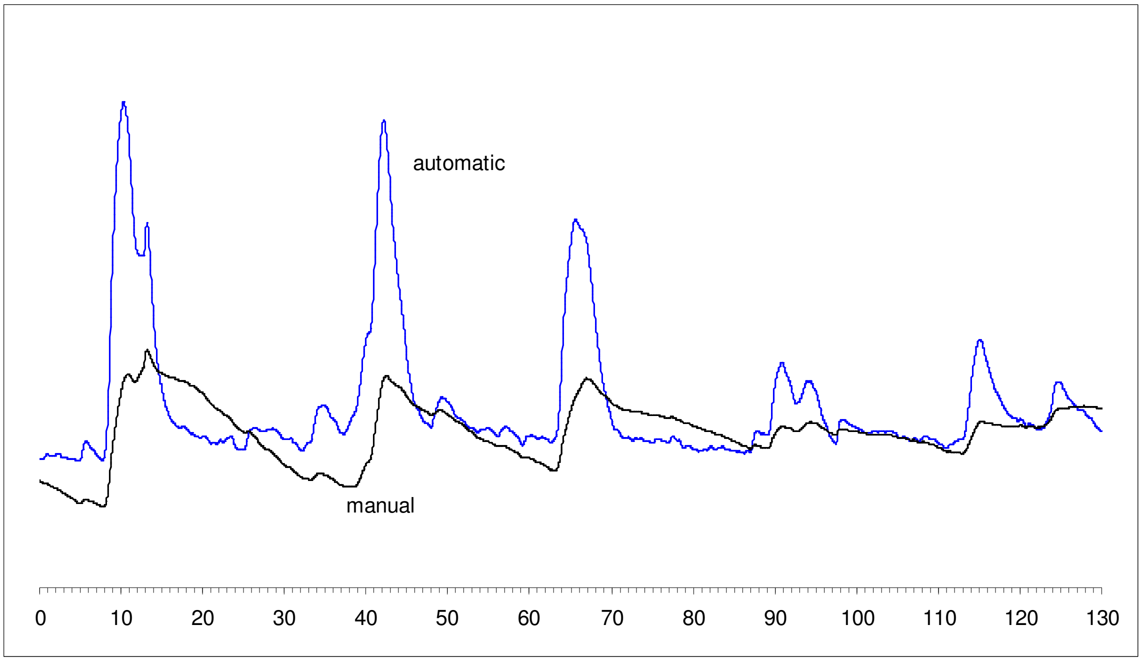


Figure A-3. Automatic and Manual Mode Electrodermal For Same Signal

Detrending

Detrending removes the long-term trends in the signals which occur over the course of the interview and are not related to a particular question. While in some cases there may be an underlying physiological cause for a trend, in polygraph this can be due entirely to a leak in the cardio-cuff, for example. This aliasing of causes prevents trends from possibly being used in our algorithm. Detrending can be accomplished in several ways. For example, a least-squares-fit line or a quadratic curve could be subtracted-off the original data. However, due to signal distortions and other irregularities, these techniques were not found to be suitable.

A better technique was found to be moving average detrending. In this technique the average of all points in an interval, centered about the point being detrended, is subtracted; and this is repeated separately for each point in the signal. Using the mean of the interval as the average is computationally efficient. However, the mean is very sensitive to any signal distortions or very large reactions. Therefore, we found it best to use the median as our average instead of the mean. Figures A-4 and A-5 shows the same signals, before and after they have been median detrended. As can be seen, the signals have been "leveled" while local changes have been retained.

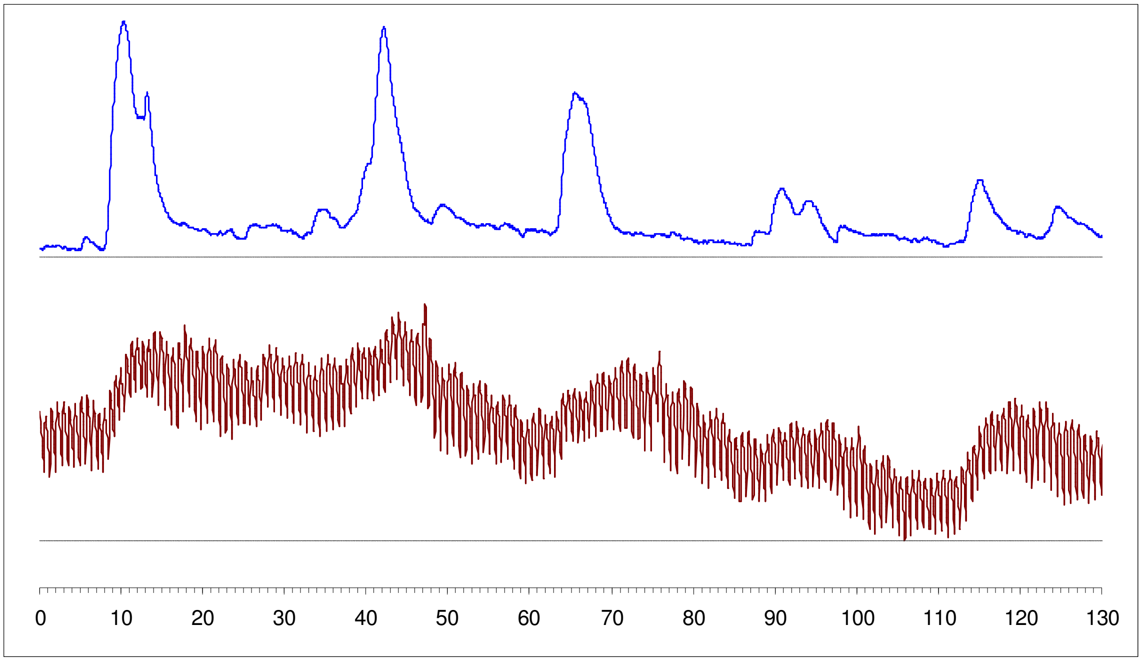


Figure A-4. Signals Before Detrending

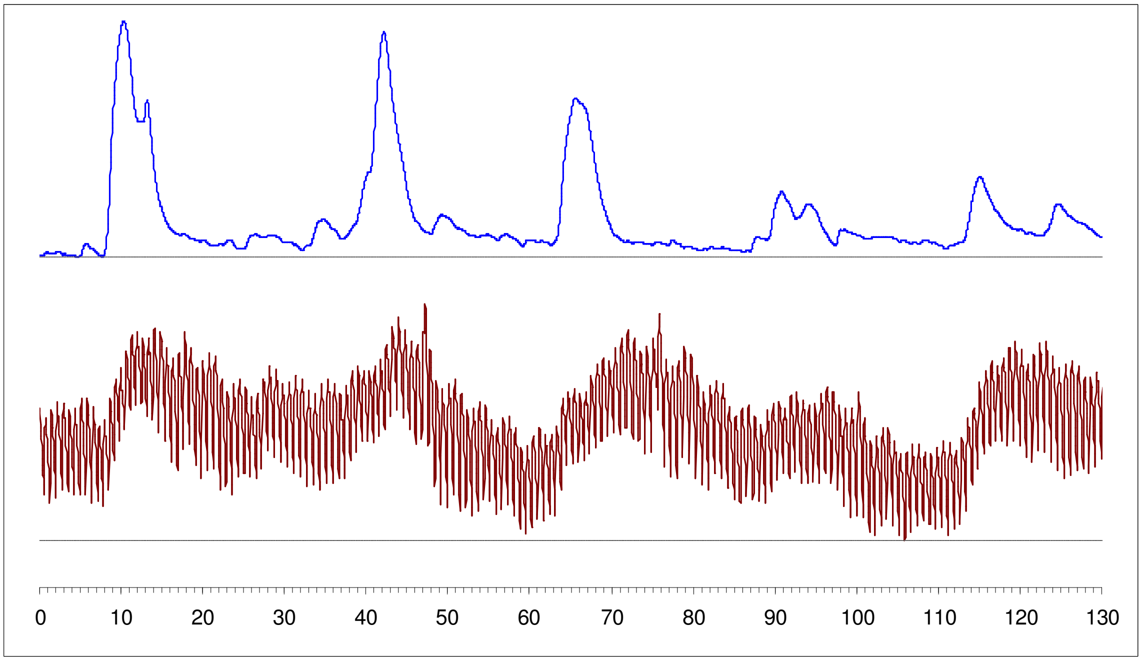


Figure A-5. Signals After Detrending

Signal Standardization

Signal standardization is a technique for removing the arbitrariness inherent in all of the signal amplitude measurements. It is accomplished by standardizing each sample point of a signal against all of the other samples taken of that signal for an entire chart, or interview in the case of the PCASS. This standardization of each signal is done using the median value of the signal and the signal's inter-quartile range. The median is subtracted from each sample point and the point is scaled by the 75th-25th percentile (inter-quartile) range. This results in a signal centered at zero and of such width that the middle 50% of the signal is one unit wide. This is depicted in Figure A-6. The reference lines show the actual 25th and 75th percentile amplitudes. Note that a signal spends 25% of its time above the 75th percentile amplitude and 25% of its time below its 25th percentile. Any apparent asymmetry, for example in the electrodermal, in Figure A-6 is caused by the magnitude of the excursions above and below these amplitudes. However, the 75th and 25th percentiles are unaffected by the sizes of these excursions and are strictly established by the band within which the signal spends 50% of its time, and not at all on the 50% of the signal outside of this band. This makes this form of standardization very robust with respect to signal distortions such as movement artifacts.

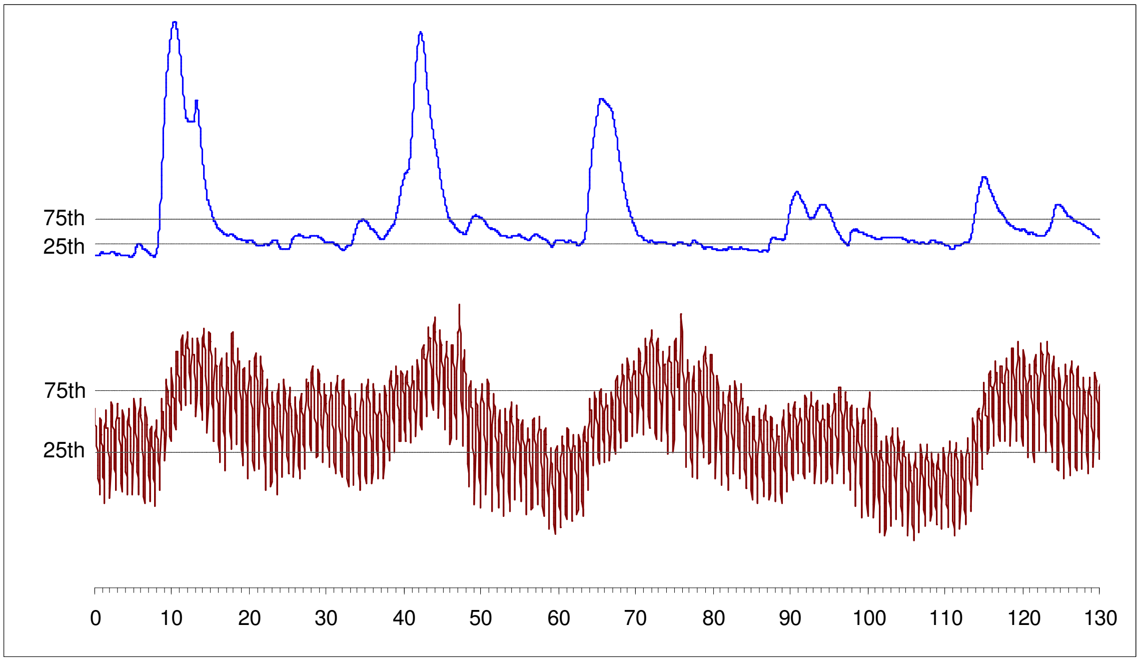


Figure A-6. Signal Standardization By Interquartile Range

Smoothed Electrodermal

The automatic mode electrodermal signal is characterized by quickly rising and falling reaction bumps. A typical reaction has one such bump. However, some reactions have several such bumps occurring in close proximity. Such reactions are called complex. One challenge with complex reactions is how to extract features—is one large reaction equivalent to a mid-sized complex reaction? The smoothed electrodermal signal was developed to transform complex reactions into more simple reactions. The blue lines in Figure A-7 illustrate a complex reaction on the left and a single bump reaction on the right. The black lines in Figure A-7 show the smoothed automatic mode electrodermal reactions. While the automatic mode reactions look quite different, the smoothed electrodermal reactions look more similar.

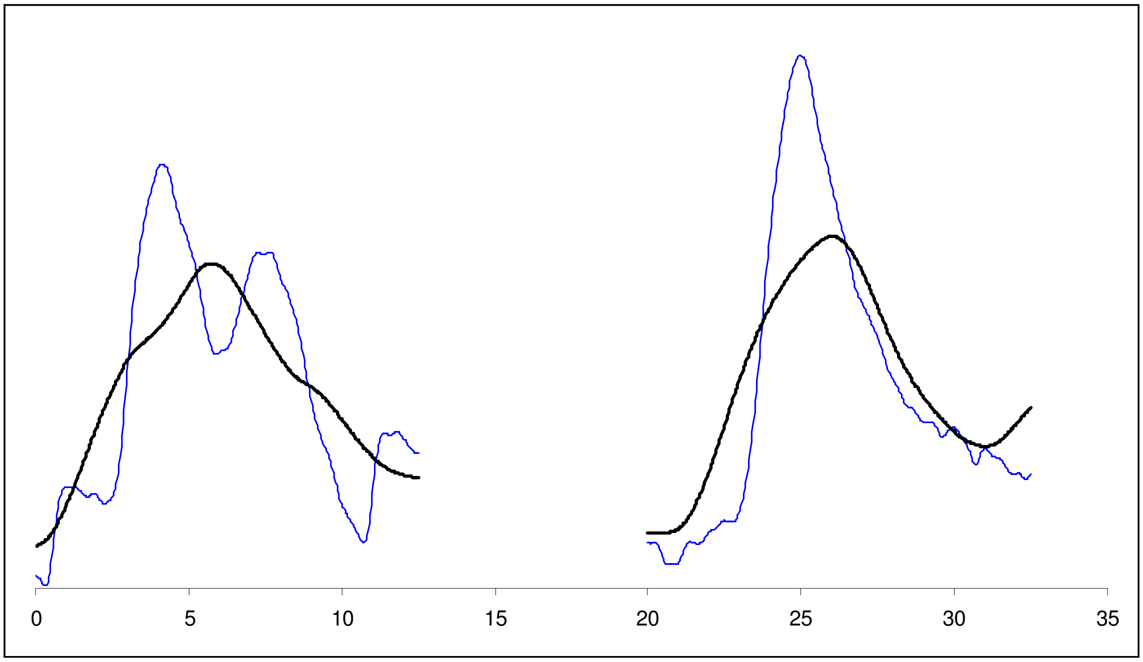


Figure A-7. Automatic Mode (blue) and Smoothed (black) Electrodermal

Cardiovascular Splitting

The cardiovascular signal collected from either the plethysmograph or the blood-pressure cuff is actually a composite measure of two distinct phenomena: the overall volume of blood and the pulsatile contraction-relaxations of the heart. These two phenomena differ greatly in their frequencies, the blood-volume increasing and decreasing over several seconds while the pulse occurs about once or twice a second. It is very useful to separate these two phenomena for analysis.

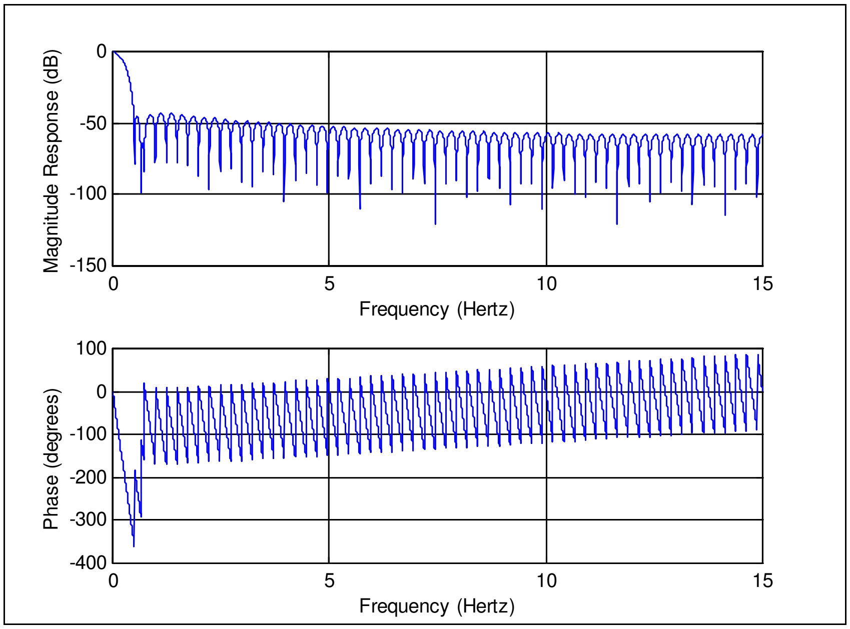


Figure A-8. Cardiovascular Low Pass Filter Characteristics

This is accomplished by digital filtering. A low-pass Finite Impulse Response (FIR) filter with the characteristics shown in Figures A-8 and A-9 is applied to the cardiovascular signal. This results in the low frequency blood-volume information being retained and the pulses being eliminated. Figure A-10 shows the original cardiovascular signal overlaid with the filtered blood-volume signal. It can be seen that when overlaid, the blood-volume signal closely follows the middle of the composite cardiovascular signal. Figure A-11 provides another view where the high frequency pulse information is also shown. The pulse signal is the residual obtained by subtracting the low frequency blood volume signal from the cardiovascular signal.

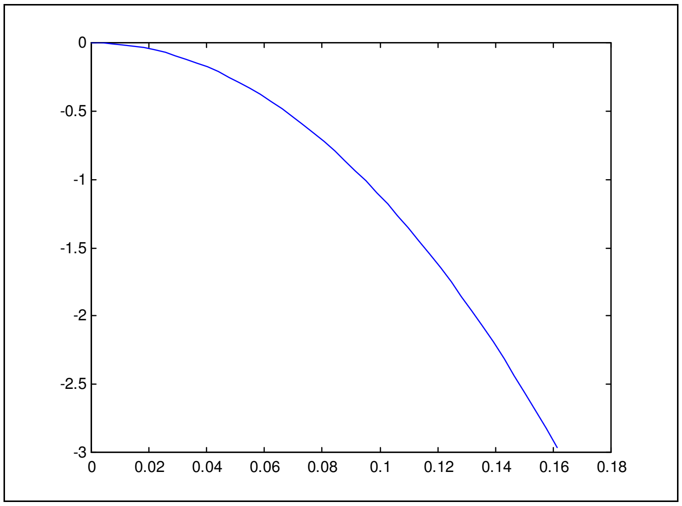


Figure A-9. Cardiovascular Low-Pass Filter Magnitude Response at Expanded Scale

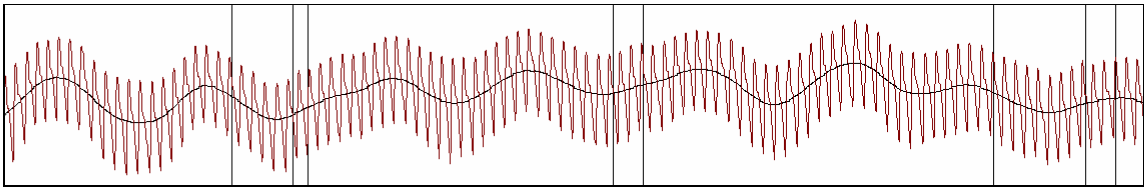


Figure A-10. Blood Volume Signal Overlaid on Cardiovascular

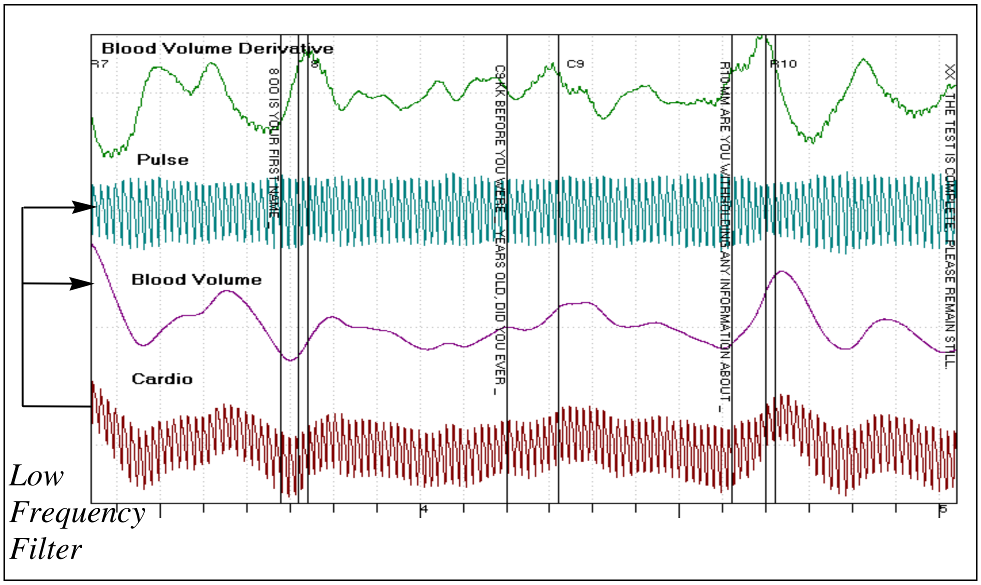


Figure A-11. Splitting Cardiovascular Signal into Pulse and Blood Volume

Cardio Tachometer

Before any signals can be derived from heartbeats, the individual heartbeats must be identified. We do this beginning with the pulse signal (see Figure A-11) and identify the local minima. We define a heartbeat as the interval between adjacent diastiolic tips. Figure A-12 illustrates a beat (red line) and its features. A beat typically begins with a sharp rise from the diastolic tip to the systolic peak followed by a slower fall or by the secondary rise/fall of the dicrotic notch. The smallness of the dicrotic notch relative to the overall beat and it's height above the diastolic tips suggest that baselining can be applied to identify the diastolic tips. Baselining begins by identifying significant local minima. Lines connecting-the-dots between these local minima are drawn as seen in the top part of Figure A-13. The baselined pulse is produced by subtracting the baseline from the pulse signal. Local minima with value 0 become the location of the diastolic tips. Cardio tach is derived from these tip-to-tip intervals.

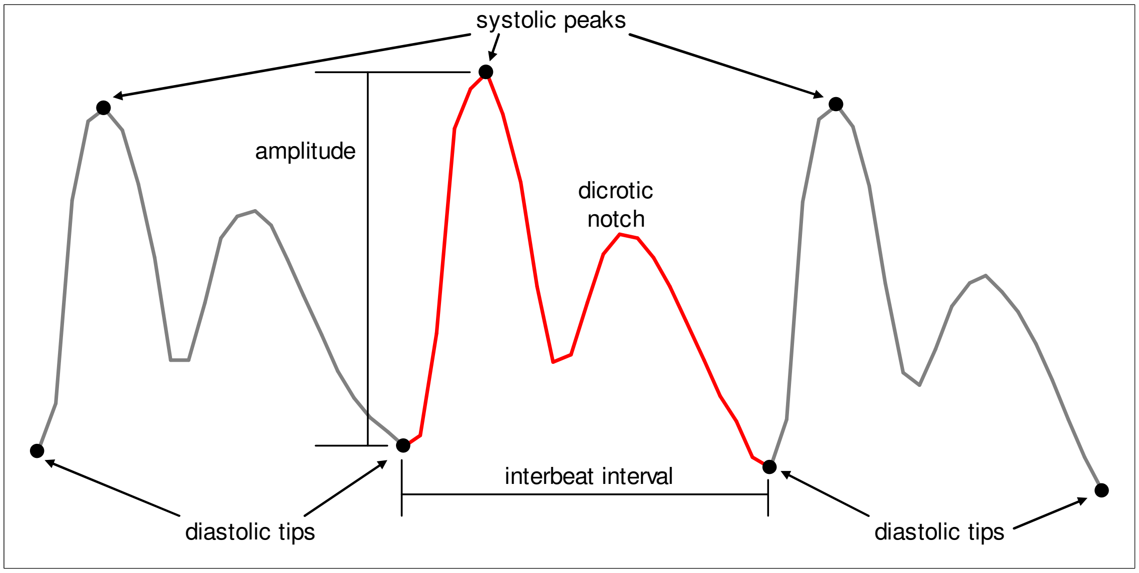


Figure A-12. Heartbeat features. Red denotes one beat

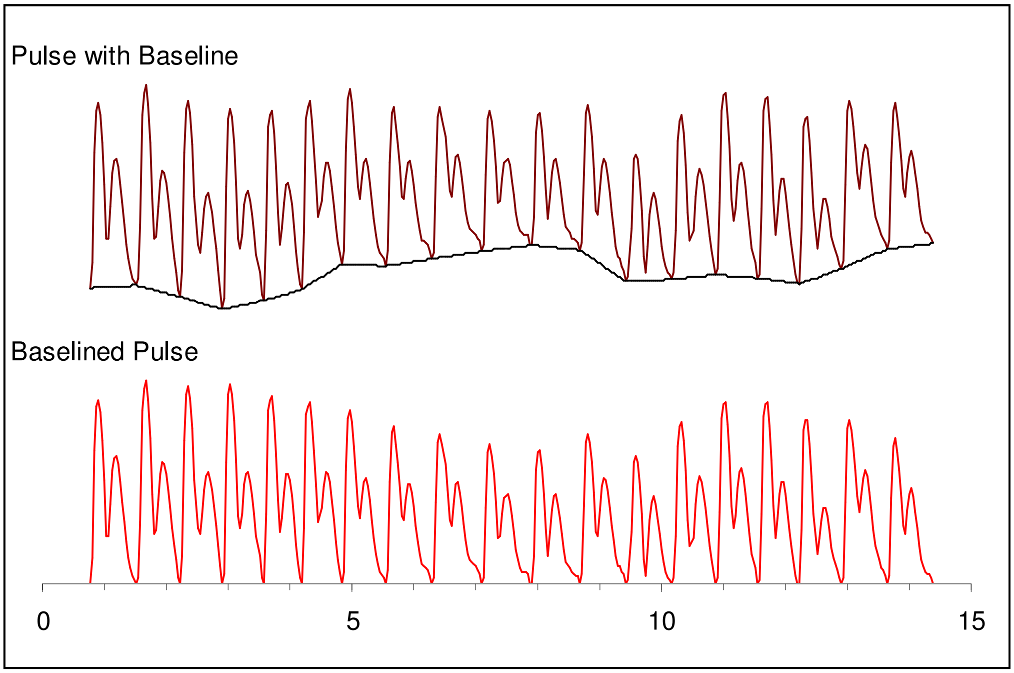


Figure A-13. Baselined pulse used to identify individual beats

The interbeat interval or heartbeat length is defined as the distance between the current diastolic tip and the next diastolic tip, call this interbeat interval distance *d*. Cardio tach is defined as –*d*. Although not strictly a frequency, experimentation showed that using the 1/*d* frequency produced a model with poor extrapolation properties. This could be due to the increased variability in 1/*d* for small *d*. However, using –*d* captures similar information as 1/*d* in that faster beats have large values and slow beats have small values. The value –*d* is assigned to all points in the beat producing a square-wave signal. The square-wave signal is smoothed with the same filter used to split the cardiovascular signal.

Our experience with other physiological signals shows that derivatives contain additional information. The *cardio tach derivative* signal is derived from the smoothed cardio tach signal using a secant method. Figure A-14 illustrates a cardio tach derivative signal (upper blue line) together with the cardio tach and pulse signal.

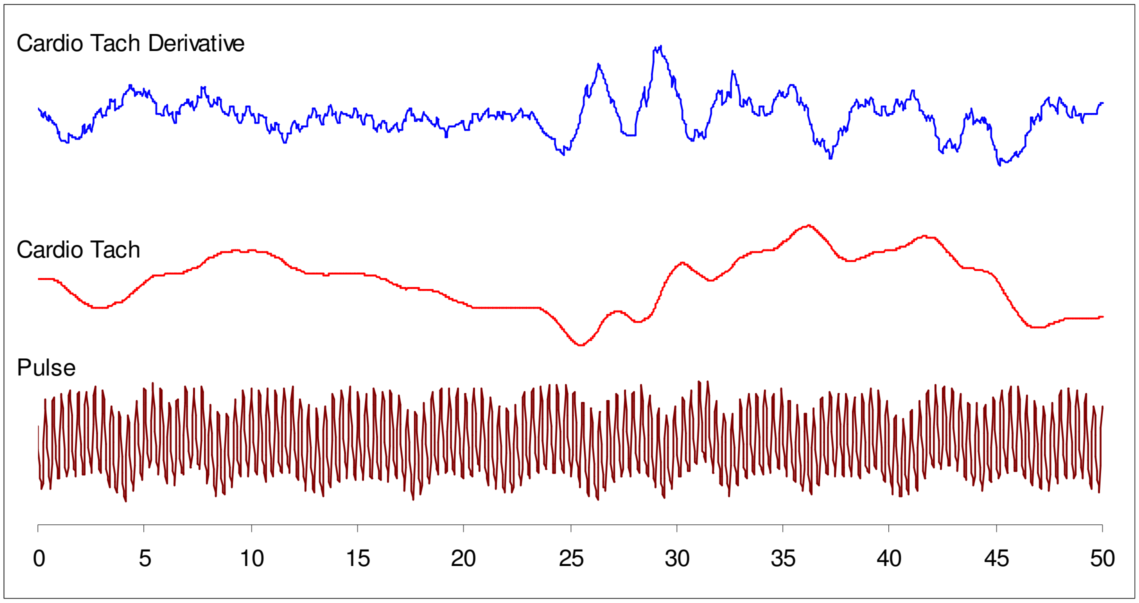


Figure A-14. Cardio tach and it's derivative

Where the heart rate slows down and beat length increases, cardio tach has smaller values. Regions of faster heart rate have smaller beat lengths and higher cardio tach values. We found that the heart rate slow-down at around 8-seconds after question onset has information and is associated with an orienting response [2]. The larger the cardio tach decrease at around 8-seconds, the stronger the indication of deception. This decrease often occurs after the peak of the automatic mode electrodermal reaction as seen in Figure A-15.

Derivatives

While the basic signals give important information by their amplitudes, how quickly these signals are changing during a reaction is also important. This information is contained in the derivatives of the signals. The derivative of a signal yields its rate of change. It provides a generalized measure of the slope of a signal. The top signal in Figure A-11 is the blood-volume derivative. It measures how sharply the overall cardiovascular signal rises during a reaction. The derivative at time point *t* is defined as $SD(t) = \left(S(t+1) - S(t-1)\right)/2$, where *SD* is the derivative and *S* is the signal.

Feature Development and Processing

Features characterize the physiological responses to questions in the various signals. Intervals of data after the beginning of the questions (onset) are used to compute the features. The features are extracted for both relevant and control questions. The relevant question features are standardized against the control question features, essentially making the control questions the same for all interviewees.

Feature Class Definitions

The shapes of the various signals in response to different questions (reactions) are characterized using features. For example, a feature might be the difference between the highest and lowest amplitudes reached by a signal during some period after a question is asked. This feature is known as the range and is shown for the electrodermal signal in Figure A-15. In general, we have identified three classes of features which are useful for characterizing reactions in terms of a reaction's amplitude, duration, latency, timing, shape, and frequency. These classes are: percentiles, time to percentiles, and percentile crossings.

Question percentile features play an important role in the algorithm. Figure A-16 shows the same reactions as Figure A-15, however they have been sorted to allow their percentiles to be calculated. The percentile features are functions of amplitude and duration: *How long was the signal at or below a given height*? For example, if a response is at or below the value ***x*** 90% of its time (the top 10% of the tracing is larger than ***x***) then the value of the 90th percentile is ***x.*** Many of the features used in our algorithms are percentile features.

The differences between percentiles can also be used as features and generalize the concept of range. Zero-crossings are a rough measure of a signal's frequency, which is obtained by counting the number of times that a signal crosses the zero line. A generalization of that concept is counting the number of times that a signal crosses an amplitude other than zero. This amplitude can be given be specified relatively by the signal's percentiles or in absolute, standardized amplitude units. Reaction latency, the time between the application of a stimulus (asking the question for polygraph) and the beginning of the response, is sometimes of interest. That concept can be generalized to measuring the time required to reach the various reaction percentiles.

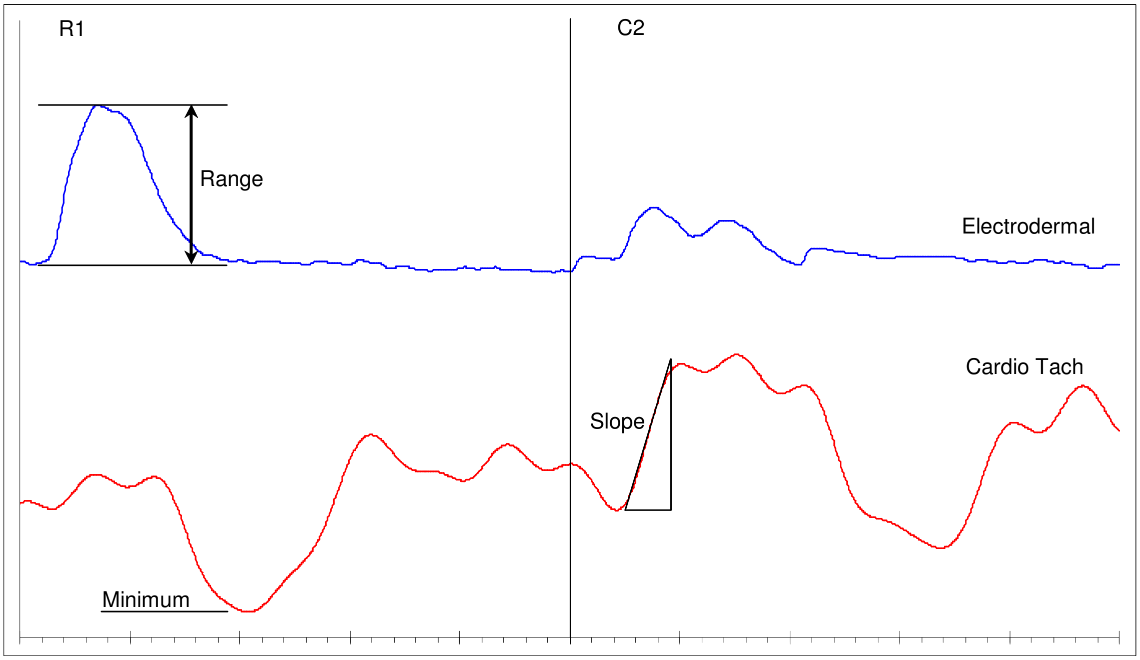


Figure A-15. Control and Relevant Reactions Showing Conceptual Features

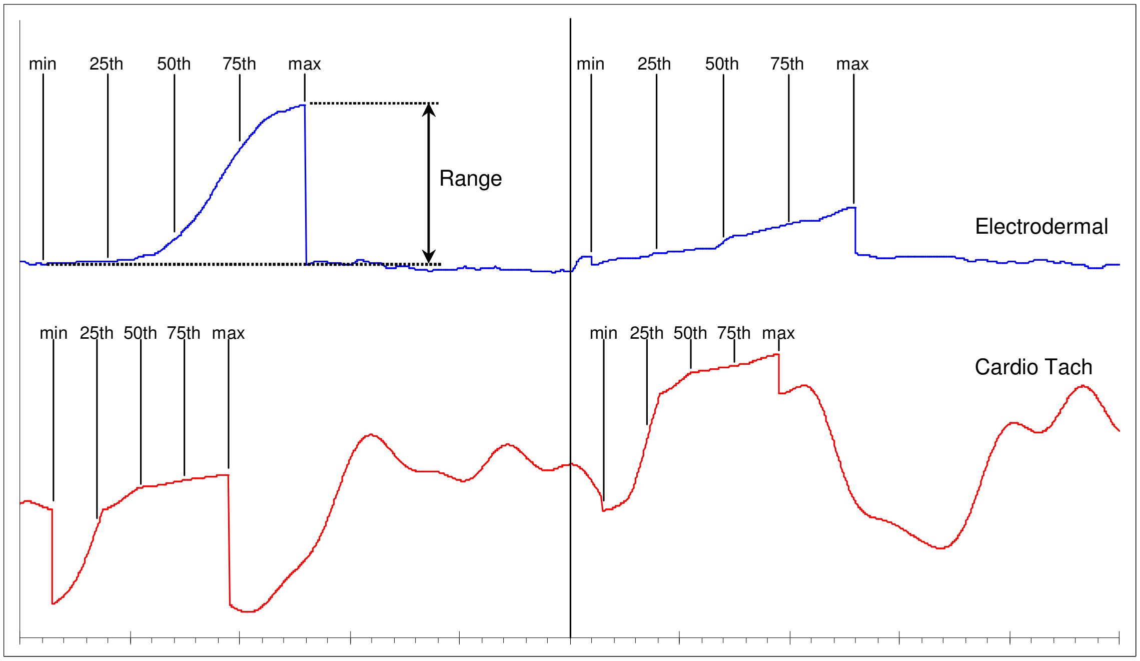


Figure A-16. Control and Relevant Reactions Showing Some Percentile Features

Response Intervals

Features are calculated for the portion of the signal which occurs as a reaction to a question. The features are extracted at each question from an interval of data defined in terms of its beginning and end after the question's onset. The end of the interval we found to be most useful does not necessarily correspond visually to the end of a reaction in a signal. The beginning and end times for each signal were identified through a series of studies as those that maximized the information in their respective important features, such as the electrodermal range and the blood volume derivative 75th percentile. Figure A-17 gives overall intervals for the various signals.

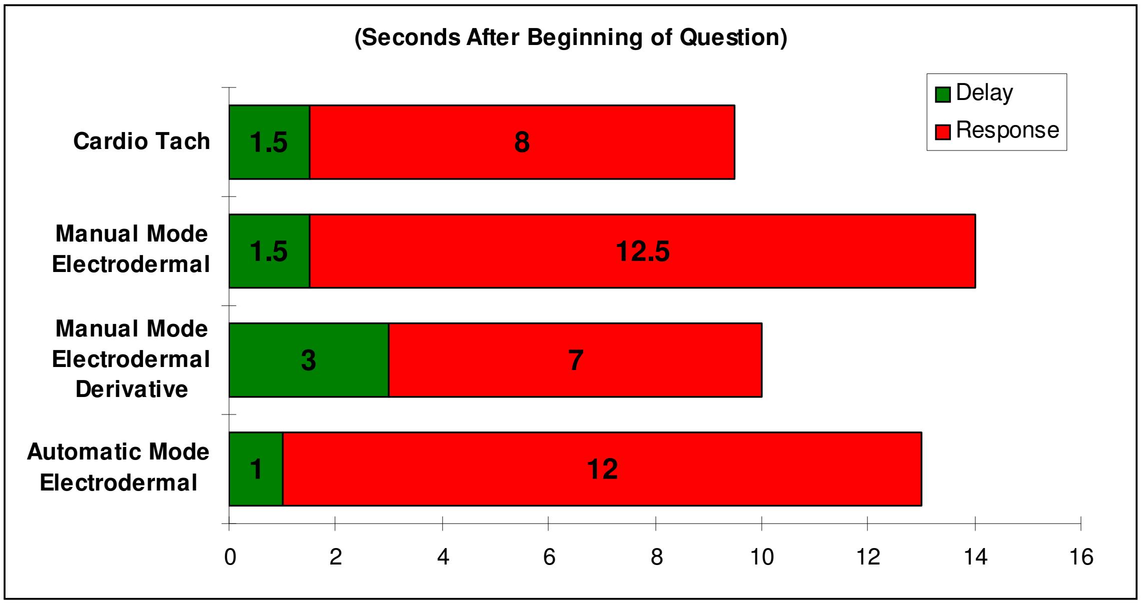


Figure A-17. Total Response Intervals

Feature Standardization

Interview evaluation is predicated on the concept of differential scoring, where an interviewee's responses to relevant questions are compared to control question responses occurring within the same examination. In our evaluation algorithms, this comparison is implicit in the process of standardizing the features. This is done as follows. Rather than using the mean for all of the questions in the standardization, just the mean of the control questions is used. This allows each relevant question to be compared to the average control. However, all of the questions are used to calculate the standard deviation. The standard deviation for the Controls is combined with that of the Relevants into a pooled standard deviation. The formula for this computation is:

$$R'_i = \frac{R_i - \overline{m}_C}{S_{CR}}$$

where R'_i is the i^{th} standardized relevant question feature,

R_i is the i^{th} relevant question feature,

$\overline{m}_C$ is the mean of the irrelevant (or control) features, and

$S^2_{CR} = \dfrac{\Sigma(R_i - \overline{m}_R)^2 + \Sigma(C_i - \overline{m}_C)^2}{(\text{number of questions} - 2)}$ is the pooled variance,

$\overline{m}_R$ is the mean of the relevant features, and

C_i is the i^{th} irrelevant (or control) question feature.

The pooled standard deviation takes advantage of consistency of reactivity (or lack thereof) for both controls and relevant questions. A small but consistent difference between control and relevant questions will be standardized differently from a small difference where one or both sets of reactions vary widely. The standardization is done separately for each feature.

Spot List

A spot is comprised of all repetitions of the same phrasing of a relevant question. The degree of overlapping semantics between various relevant questions is what separates single-issue from multiple issue interviews. Even when deceptive to more than one issue, it is not uncommon for an interviewee to focus on a particular question. Spot evaluation allows the issue of greatest concern to be both identified and measured. For this purpose a distinct spot model is built which draws from the same standardized relevant question feature set as for the question features, except that the features for each spot are averaged.

The overall spot is a spot assuming that all relevant question phrasings are equivalent, e.g.:

- Did you steal the car?
- Did you steal the car missing from John's driveway?.

Each interview has only one overall spot consisting of all of the relevant questions asked. Features for the overall spot are averaged over all questions.

Naïve Bayes Classifier

Consider a training dataset with n_1 observations from population 1 and n_2 observations from population 2. Let X_{ij} be the jth variable (j=1,…,K) in the ith observation (i=1,…, n_1) from population 1. Let Y_{ij} be the jth variable (j=1,…,K) in the ith observation (i=1,…, n_2) from population 2. First, compute sample means and standard deviations for the two samples. For sample 1, let

$$\bar{X}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{ij} \text{ and } \mathrm{sx}_j = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{ij} - \bar{X}_j)^2}$$

be the sample mean and standard deviation, respectively, for the jth variable in sample 1. Sample 1 statistics estimate population 1 parameters. For sample 2, let

$$\bar{Y}_j = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{ij} \text{ and } sy_j = \sqrt{\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_{ij} - \bar{Y}_j)^2}$$

be the sample mean and standard deviation, respectively, for the jth variable in sample 2.

The second part of the classification is the density function. For PCASS, we focused on Cauchy densities. A Cauchy density is a bell-shaped curve with density function

$$f(z) = \frac{1}{\pi\left(1 + z^2\right)}.$$

Let $x = a + bz$, where $b \neq 0$ and z has a cauchy density. Then the density of x is

$$f(x;a,b) = \frac{1}{\pi b\left(1 + \left(\frac{x-a}{b}\right)^2\right)}.$$

All variables and observations use the same family of densities either all normal or all Cauchy.

To classify a new observation o, compute a sample 1 density $f_{1j}(o;\bar{X}_j, sx_j)$ and a sample 2 density $f_{2j}(o;\bar{Y}_j, sy_j)$ for each variable j. Define the population 1 probability as

$$p_1 = \prod_{j=1}^{K} f_{1j}(o;\bar{X}_j, sx_j)$$

and the population 2 probability as

$$p_2 = \prod_{j=1}^{K} f_{2j}(o;\bar{Y}_j, sy_j).$$

The naïve Bayes classification probability that observation o belongs to population 1 is defined as $P(\text{observation belongs to population 1}) = \frac{p_1}{p_1 + p_2}$. When constructing the decision rules, naïve Bayes classifications probabilities are computed for each training observation.

Scoring Spots with the Spot Model

Our decision rule uses both the maximum spot and the overall spot probabilities. The maximum spot is that spot with the largest probability of deception. The overall spot is considered a separate item and is not used in computing the maximum spot. When using maximum spot we observed that the maximum probability was heavily skewed towards 1. In order to reduce the skew in the probabilities, all spots, including the overall spot, are rescaled. First the probabilities are converted to a score using the logit transformation. The scores are rescaled and converted back to probabilities. Scores were computed for the maximum spot for all truthful and deceptive interviews in the training data set. An average truthful score and average deceptive score were computed. Scores were rescaled such that the average truthful max spot had a value of -2.5 and the average deceptive max spot score had a value of 2.5.

The first step in scoring the spots is to assign the Naïve Bayes probabilities p to each spot. Then the maximum spot is determined along with its probability p_{max}. Then the Naïve Bayes probability p_{over} is assigned to the overall spot. These two probabilities are converted to scores as follows:

```
    if prob<0.0000000000001 then prob = 0.0000000000001,
    if prob>0.9999999999999 then prob = 0.9999999999999,
    score = log(prob/(1-prob)).
```

The if statements prevent computer overflow/underflow problems. The last step being the logit transformation. The scores are rescaled by the linear transformation

```
    score = (score - 1.2488646274861426) / 0.86788975651495193.
```

Finally the rescaled scores are converted back to probabilities by

```
  prob = 1.0 / (1.0 + exp(-score)).
```

Outcome Decision Rules

Determining the final outcome decision rule is somewhat difficult. The outcome decision rule depends on both the maximum and overall rescaled spot probabilities and it produces one of three possible outcomes. There is a tradeoff between more inconclusive outcomes and overall accuracy. There is also a tradeoff between false positive and false negative outcomes. We explored a wide range of possible decision rules and selected one which agreed best with our sponsor's desire to bound the overall inconclusive rate at 30% and to minimize false negative errors. We employed several statistical tests to assess overtraining. These include comparisons of the means and standard deviations of the training and test data set scores (see above) as well as 2×3 contingency table analysis of the color outcomes for the training and test data sets. As seen in Figure 13 we adopted the following decision rules: By default, an interview's outcome is Yellow. If the maximum spot probability ≤ 0.34 and the overall spot probability ≤ 0.03 then the outcome is Green. If maximum spot probability ≥ 0.63 and the overall spot probability ≥ 0.26 then the outcome is Red. The outcome is also Red if the maximum spot probability ≥ 0.93 or overall spot probability ≥ 0.40.

References

1. Boucsein, Wolfram, Electrodermal Activity, Plenum, 1992
2. Hugdahl, Kenneth, Psychophysiolgy: the Mind-Body Perspective, Harvard University Press, 1995
3. Swinford, J, Manually Scoring Polygraph Charts Utilizing the Seven-Position Numerical Analysis Scale at the Department of Defense Polygraph Institute, Polygraph, Vol. 28, No. 1, 1999
4. Matte, J. A., Forensic Psychophysiology Using a Polygraph, J. A. M. Publications, Williamsville, NY
5. Metz, CE., ROCKIT IBM compatible version 0.9B. Department of Radiology, The University of Chicago, March 1998
6. Swets JA, Pickett RM., Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York: Academic Press, 1982
7. Hosmer, Davaid W. and Stanley Lemesho, Applied Logistic Regression, Wiley, New York, 1989
8. D. E. Olsen, J. C. Harris, M. H. Capps, N. Ansley, “Computerized Polygraph Scoring System”, Journal of Forensic Sciences, Vol. 42, No.1, January 1997
9. Haykin, Simon, Neural Networks, A Comprehensive Foundation, Prentice Hall, New Jersey, 1999
10. McLachlan, G., Discriminant Analysis and Statistical Pattern Recognition, John Wiley and Sons, 1992
11. Witten, Ian and Eibe Frank, Data Mining, Morgan Kaufmann, New York, 2000
12. Duda, Richard, Peter Hart and David Stark, Pattern Classification, John Wiley and Sons, New York, 2001