Evaluating Polygraph Data

Aleksandra Slavkovic
sesa@stat.cmu.edu

Abstract

The objective of automated scoring algorithms for polygraph data is to create reliable and statistically valid classification schemes minimizing both false positive and false negative rates. With increasing computing power and well developed statistical methods for modeling and classification we often launch analyses without much consideration for the quality of the datasets and the underlying assumptions of the data collection. In this paper we try to assess the validity of logistic regression when faced with a highly variable but small dataset.

We evaluate 149 real-life specific incident polygraph cases and review current automated scoring algorithms. The data exhibit enormous variability in the subject of investigation, format, structure, and administration making them hard to standardize within an individual and across individuals. This makes it difficult to develop generalizable statistical procedures. We outline steps and detailed decisions required for the conversion of continuous polygraph readings into a set of features. With a relativelly simple approach we obtain accuracy rates comparable to those currently reported by other algorithms and manual scoring. Complexity that underlines assessment and classification of examinee's deceptiveness is evident in a number of models that account for different predictors yet give similar results, typically "overfitting" with the increasing number of features. While computerized systems have the potential to reduce examineer variability and bias, the evidence that they have achieved this potential is meager at best.

1 Introduction

William M. Marston, a psychologist and the Wonder Woman creator, had an idea for an infallible instrument for lie detection some 25 years before he created this heroine whose magic lasso forces villains to tell the truth. In 1915 while a graduate student at Harvard University he reported that blood pressure goes up when people lie. This paved the way for the invention of the polygraph, a 'lie detector' machine, in the early 1920s by J. A. Larson and L. Keeler[1, 23].

Polygraphs are used by law enforcement agencies and the legal community for criminal investigations, in the private sector for pre-employment screening, and for testing for espionage and sabotage. Polygraph proponents claim high accuracy rates of 98% for guilty subjects and 82% for innocent [2, 26]. These rates are typically calculated by leaving out inconclusive cases and ignoring the issue of sampling bias, e.g. when the accuracies and inter-raters reliability are calculated using only subjects for which there is an independent validation of their guilt or innocence (i.e., ground truth).

The polygraph as an instrument has been recording changes in people's relative blood pressure, respiration and the electrodermal response (palmar sweating or galvanic skin response) in some form since 1926. These psychophysiological responses, believed to be controlled by the autonomic nervous system, are still the main source of information from which the polygraph examiners deduce an examinee's deceptive or non-deceptive status. The underlying premise is that an examinee will involuntarily exhibit fight-or-flight reactions in response to the asked questions. The autonomic nervous system will, in most cases, increase the person's blood pressure and sweating, and affect the breathing rate. These physiological data are evaluated by the polygraph examiner using a specified numerical scoring system and/or statistically automated scoring algorithms. The latter are the main focus of this report. Current methods of psychophysiological detection of deception (PDD) are based on years of empirical work, and often are criticized for a lack of thorough scientific inquiry and methodology.

The objective of automated scoring algorithms for polygraph data is to create reliable and statistically valid classification schemes minimizing both false positive and false negative rates. The statistical methods used in classification models are well developed, but to the author's knowledge, their validity in the polygraph context has not been established. We briefly describe the polygraph examination framework and review two automated scoring algorithms relying on different statistical procedures in Section 2. Section 3 provides some background on statistical models one might naturally use in settings such as automated polygraph scoring. In Section 4 we evaluate collection of real-life polygraph data of known deceptive and nondeceptive subjects. In Section 5 we outline steps and detailed decisions required for the conversion of continuous polygraph readings into a set of numeric predictor variables and present results of a logistic regression classifier. Our approach is simpler than other proposed methods, but appears to yield similar results. Various data issues that are not addressed or captured by the current algorithms indicate a deficiency in validity of methods applied in the polygraph setting.

2 Background

2.1 The Polygraph Examination

The polygraph examination claims to measure psychophysiological detection of deception (PDD). It measures some emotional responses to a series of questions from which deception is inferred. The PDD examination can be divided into three parts: pre-test, in-test, and post-test. During the *pre-test* the examiner explains to the examinee the theory of the test and the instrument, and fomulates the test questions. Experienced polygraphers believe that results may be erroneous if this part is not conducted properly. An "acquaintance" test is run to demonstrate that the examinee's reactions are captured. During the *in-test* phase, a series of 8 to 12 Yes and No questions are asked and data are collected. Typically, the same series of questions are asked at least 3 times, with or without varying the order of the questions. One repetition represents a polygraph chart or test limited to approximately 5 minutes. In the *post-test* the examinee's responses are reviewed with the examinee; the examiner often obtains a confession from guilty subjects (sometimes from innocent ones too!). If the examinee provides further explanation regarding his/her responses, then new questions may be composed and the test repeated [23, 7].

There are three main types of questions. *Irrelevant* or *neutral* questions such as "Is today Tuesday?" are meant to stabilize the person's responses with respect to external stimuli such as the examiner's voice. *Relevant* questions address the main focus of the examination: "Did you steal that money?" *Comparison*, also known as *control*, questions address issues similar in nature but unrelated to the main focus of the exam. "Have you ever cheated a friend out of anything?" is an example of a 'probable-lie' control question where the examiner is not sure if the subject is lying or not. For a 'direct-lie' control question, the examiner instructs the examinee to be deceptive by thinking or imagining an uncomfortable personal incident. Since the topic of each examination varies, the questions are custom-made for the examinee.

Evaluation of the charts is performed by relative comparison of the responses on the relevant and control questions. It is expected that the deceptive person will show stronger reactions to the relevant questions, while the innocent person will be more concerned with control questions and show stronger reactions on those than on relevant questions. Depending on the test format, questions appear in different positions in the question sequence.

2.2 Type of Test

There are two broad categories of PDD: specific issue and screening. *Specific* issue examinations address specific known events that have occurred (robbery, theft, rape, murder). Each of these PDD categories has a number of different test formats; most are versions of the Control Question Test (CQT). The formats differ by the number, type, and order of questions asked, and in the way the pre-test is conducted. We are concerned with two CQT formats: the Zone Comparison Test (ZCT), and the Multiple General Question Test (MGQT). Each of these may have variations depending on the agency that conducts the examination and examiner's training. According to the Department of Defense Polygraph Institute (DoDPI), the ZCT usually has the same number of control(C) and relevant(R) questions. A typical sequence is CRCRCR, allowing for comparison of the nearest relevant and control questions on all three charts. The MGQT has more relevant than control questions. A proposed sequence is RRCRRC, but this is not always the case in practice. These sequences may be interspersed with irrelevant(I) questions. The first two charts usually have the same question order. The third chart looks more like the ZCT format. Both ZCT and MGQT may have other "wild-type" questions introduced during the exam [7]. The order of questions may or may not be randomized across different charts. All these elements add variability to the data. An example set of questions for ZCT and MGQT formats is in Table A1 of the Appendix.

Screening tests address events that may have occurred (espionage, sabotage, disclosure of classified information). The most commonly used is the Test of Espionage and Sabotage (TES) which asks all individuals the same type of questions in the same order on two charts. An example of TES questions is presented in Table A2.

2.3 Instrumentation and Measurements

A polygraph instrument records and filters the original analog signal. The output is a digital signal, a discretized time series with possibly varying sampling rates across instruments and channels. The polygraph typically records thoracic and abdominal respirations, electrodermal and cardiovascular signals (Figure 1).

Pneumographs positioned around the chest and the abdomen measure the rate and depth of respiration. Subjects can control their breathing and influence the recorded measurements. Changes in respiration can also affect heart rate and electrodermal activity. For example coughing is manifested in electrodermal activity.

Electrodermal activity (EDR) or sweating is measured via electrodes (metal plates) placed on two fingers, and it is considered the most valuable measure in lie detection. When a small current is passed through the skin, skin conductance (SC) or its reciprocal, skin resistance (SR)is recorded. The autonomic nervous system in a stressful situation typically increases secretion of the eccrine glands, and thus lowers the resistance. When SC is measured, the readouts can be directly interpreted as the person's reactions to the questions. When SR is measured, however, there needs to be an adjustment with respect to the basal level in order to assess the EDR activity. Some have argued that the size of the response to a question depends on which of SC or SR is recorded [9]. This is a controversial issue discussed in more detail in the psychophysiological literature and in [7].

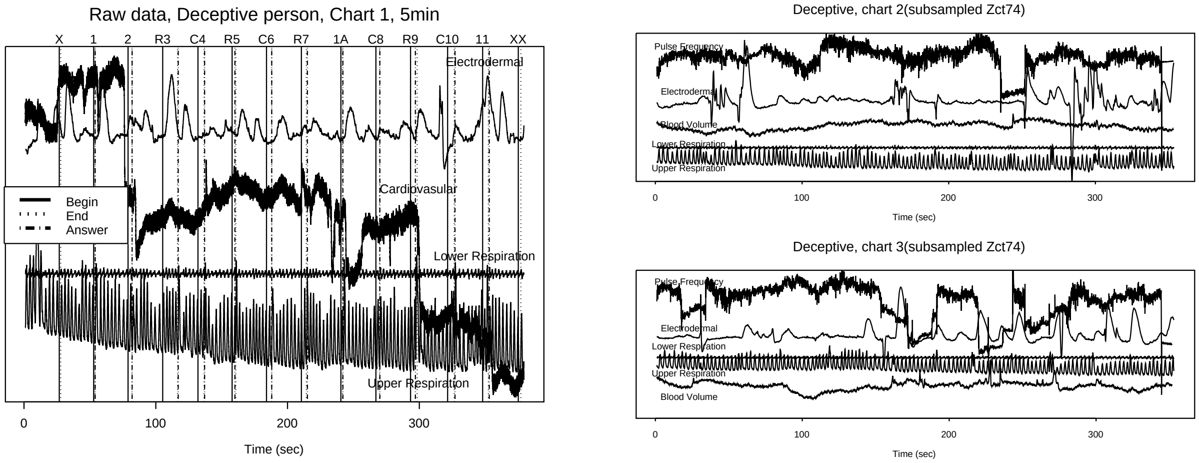


Figure 1: The figure on the left is the raw data from a deceptive person on chart 1. The lower two recordings are thoracic and abdominal respiration, the middle time series is cardiovascular signal and the upper is electrodermal signal. The scale is arbitrary. The labels on the upper axis correspond to a question sequence. The figures on the right represent data on the second and third chart of the same person. The cardiovascular signal is split into relative blood volume and pulse frequency signals as described in Section 5.1.

Cardiovascular activity is measured by a blood pressure cuff positioned above the biceps. As a hybrid signal of relative blood pressure and heart rate, it is the most complex of the four recorded measurements. The cardiovascular response is either coactivated or coinhibited by other physiological responses, making its evaluation more difficult, e.g. [0.12Hz-0.4Hz] frequency band in the heart rate is due to respiration. This coupling may differ within a person and across different environmental settings [5].

It is unclear whether these physiological responses reflect a single psychological process (such as arousal) or the extent to which they are consistent across individuals. The psychophysiological literature includes contradictory claims on how internal emotional states are mapped to physiological states, and the extent to which emotional states represent deception [7, 12, 16, 19].

2.4 Chart Evaluations

A critical part of polygraph examination is the analysis and interpretation of the physiological data recorded on polygraph charts. Polygraph examiners rely on their subjective global evaluation of the charts, numerical methods and/or computerized algorithms for chart scoring.

2.4.1 Numerical Scoring

The scoring procedure may differ by the PDD examination type, policy of the agency, and the examiner's training and experience. The 7-Position Numerical Analysis Scale scoring relies on *spot analysis* where each relevant question has a location ("spot"). The examiner looks for changes in the baseline, amplitude, duration, and frequency of the recorded signals at each spot and compares them to the activities at the nearest control question (often the strongest control is chosen). Values on a 7-point scale (-3 to 3) are assigned to the differential of the two responses. The *overall score* for each spot is calculated by summing the assigned values across charts for each channel (two respiratory tracings are added together). The *grand total* is the sum of all spot totals. The negative values indicate higher reaction on the relevant questions and the positive values indicate higher response on the control questions. A grand total score of +6 and greater indicates nondeception, -6 and less deception, and anything in between is considered an inconclusive result. These cutoffs may vary [23, 28].

2.4.2 Computerized Scoring Algorithms

Two computerized polygraph systems are currently used in connection with U.S. distributed polygraph equipment, but other systems have been developed more recently. The Stoelting polygraph instrument uses the Computerized Polygraph System (CPS) developed by Scientific Assessment Technologies based on research conducted at the psychology laboratory at the University of Utah [21, 22, 6]. The Axciton and Lafayette instruments use the PolyScore algorithms developed at the Johns Hopkins University Applied Physics Laboratory [13, 14, 25]. Three new polygraph scoring algorithms are introduced for use with the Axciton polygraph instrument: AXCON and Chart Analysis by Axciton Systems, Inc. and Identifi by Olympia. Performance of these algorithms on an independent set of 97 selected confirmed criminal cases was compared by [10](see Table A3). CPS performed equally well on detection of both innocent and guilty subjects while the other algorithms were better at detecting deceptives. Unfortunately, the method of selecting these cases makes it difficult to interpret the reported rates of misclassification. More details on the actual polygraph instruments and hardware issues and some of the history of the development of computerized algorithms can be found in [1, 23, 22].

The description here focuses on the PolyScore and CPS scoring algorithms since no information is publicly available on statistical methods utilized by these more recently developed algorithms. The methods used to develop the two computer-based scoring algorithms both fit within the general statistical framework as described below. They take the digitized polygraph signals and output estimated probabilities of deception. While PolyScore uses logistic regression or neural networks to estimate the probability of deception from an examination, CPS uses standard discriminant analysis and a naive Bayesian probability calculation to estimate the probability of deception (a proper Bayesian calculation would be far more elaborate and might produce markedly different results). They both assume equal a priori probabilities of being truthful and deceptive. The biggest differences that we can discern between them are the data they use as input, their approaches to feature development and selection, and the efforts that they have made at model validation and assessment.

PolyScore was developed on real criminal cases. CAPS, an earlier version of CPS, was developed on mock crimes, while the more recent versions rely on actual criminal cases. CAPS ground truth came from independent blind evaluations, while PolyScore relied on a mix of blind evaluations and confessions. Both algorithms do some initial data transformation of the raw signals. PolyScore uses more initial data editing tools such as detrending, filtering and baselining, while CPS tries to retain much of the raw signal. They both standardize signals, although using different procedures. They extract different features, and they seem to use different criteria to find where the maximal amounts of discriminatory information lies. Both, however, give the most weight to the electrodermal channel.

PolyScore combines all three charts into one single examination record and considers reactivities across all possible pairs of control and relevant questions. CAPS compares adjacent control and relevant questions as is done in manual scoring, but it also uses the difference of averaged standardized responses on the control and relevant questions to discriminate between deceptive and non-deceptive people. CPS does not have an automatic procedure for the detection of artifacts, but it allows examiners to edit the charts themselves before the algorithm calculates the probability of truthfulness. PolyScore has algorithms for detection and removal of artifacts and outliers, but it claims that the specific details are proprietary and will not share them.

A more detailed review of the computerized scoring systems can be found in Appendix F of [7]. Computerized systems have the potential to reduce bias in the reading of charts and inter-rater variability. Whether they can actually improve accuracy also depends on how one views the appropriateness of using other knowledge available to examiners, such as demographic information, historical background of the subject, behavioral observations.

3 Statistical Models for Classification and Prediction

This section provides some background on the statistical models that one might naturally use in settings such as automated polygraph scoring. The statistical methods for classification and prediction most often involve structure:

$$\text{response variable} = g(\text{predictor variables, parameters, random noise}), \tag{1}$$

where g is some function. For classification problems it is customary to represent the reponse as an indicator variable, y, such that $y = 1$ if a subject is deceptive, and $y = 0$ if the subject is not. Typically we estimate y conditional on the predictor variables, X, and the functional form, g. For linear logistic regression models, with k predictor variables $x = (x_1, x_2, ..., x_k)$, we estimate the function g in equation (1) using a linear combination of the k predictors:

$$\text{score}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_k x_k \tag{2}$$

and we take the response of interest to be:

$$\Pr(\text{deception}|x) = \Pr(y|x) = \frac{e^{\text{score}(x)}}{1 + e^{\text{score}(x)}}. \tag{3}$$

This is technically similar to choosing $g = \text{score}(x)$, except that the random noise in equation (1) is now associated with the probability distribution for y in equation (3), which is usually taken to be Bernoulli. We are using an estimate of the score equation (2) as a hyperplane to separate the observations into two groups, deceptives and nondeceptives. Model estimates do well (e.g., have low errors of misclassification) if there is real separation between the two groups.

Model development and estimation for such prediction/classification models involve a number of steps:

1. Specifying the list of possible predictor variables (features of the data) to be used.
2. Choosing the functional form g in model (1) and the link function.
3. Selecting the actual features from the feature space to be used for classification.
4. Fitting the model to data to estimate empirically the prediction equation to be used in practice.
5. Validating the fitted model through some form of cross-validation.

Different methods of fitting and specification emphasize different features of the data. The standard linear discriminant analysis assumes that the distributions of the predictors for both the deceptive group and the nondeceptive group are multivariate normal, with equal covariance matrices (an assumption that can be relaxed). This gives substantial weight to observations far from the region of concern for separating the observations into two groups. Logistic regression models, on the other hand, make no assumptions about the distribution of the predictors. The maximum likelihood methods typically used for their estimation put heavy emphasis on observations close to the boundary between the two sets of observations. Common experience with empirical logistic regression and other prediction models is that with a large number of predictor variables we can fit a model to the data (using steps 1 through 4) that completely separates the two groups of observations. However, once we implement step 5 we often learn that the achieved separation is illusory. Thus many empirical approaches build cross-validation directly into the fitting process, and set aside a separate part of the data for final testing.

Hastie et al.[17] is a good source of classification/prediction models, cross-validation, and related statistical methodologies. Algorithmic approachs to prediction from the data-mining focus less on the specification of formal models and treat the function g in equation (1) more as a black box that produces predictions. Among the tools used to specify the black box are regression and classification trees, neural networks, and support vector machines. These still involve finding separators for the observations, and no matter which method one chooses to use, all five of the steps listed above still require considerable care.

4 The Data

The Department of Defense Polygraph Institute (DoDPI) provided data from 170 specific incident cases that vary by the collection agency, type of crime, test formats and questions. In this report we analyzed 149 cases[1], a mix of ZCT and MGQT test formats (see Table 1). We had to discard 21 cases due to missing information on one or more charts. The type of data missing could be any combination of type of questions, onset of the questions, time of the answer and others. All data were collected with Axciton polygraph instruments.

	Deceptive	NonDeceptive	Total
ZCT	27	24	51 (51)
MGQT	90	29	119 (98)
Total	117 (98)	53 (51)	170 (149)

Table 1: Number of specific incident cases by test type and ground truth. Numbers in parenthesis are the numbers of cases used in our analysis.

Each examination (subject/case) had three to five text data files corresponding to the exam charts, each approximately five minutes long. The data were converted to text from their native proprietary format by the "Reformat" developed by JHUAPL, and the identifiers were removed. Each data file contained the sequence of questions asked during the exam, and the following fields:

1. Sample: index of observations; sampling rate is 60Hz for all measurements.
2. Time: the time, relative to the test beginning, when the sample was taken
3. Pn1: recording of the thorax respiration sensor.
4. Pn2: recording of the abdomen sensor.

[1] These data overlap with those used in the development of PolyScore

5. Edr: data from an electrodermal sensor.
6. Cardio: data from the blood pressure cuff (60 and 70 mmHg inflated).
7. Event: markings for the length of asking of the question and time of the answers.

Figure 1 shows raw data for one subject. Each time series is one of the four biological signals plus unknown error. In our analyses we use an additional series (pulse frequency that we extracted from the cardiovascular signal; see Section 5.1). Demographic data such as gender, age, and education are typically available to the examiner during the pre-test phase. We had limited demographic data and did not utilize it in the current analysis.

Respiratory tracing consists of inhalation and exhalation strokes. In manual scoring the examiner looks for visual changes in the tracings for breathing rate, baseline and amplitude, where for example 0.75 inches is the desired amplitude of respiratory activity. Upper and lower respiration recordings are highly correlated as are the features we extract on these recordings.

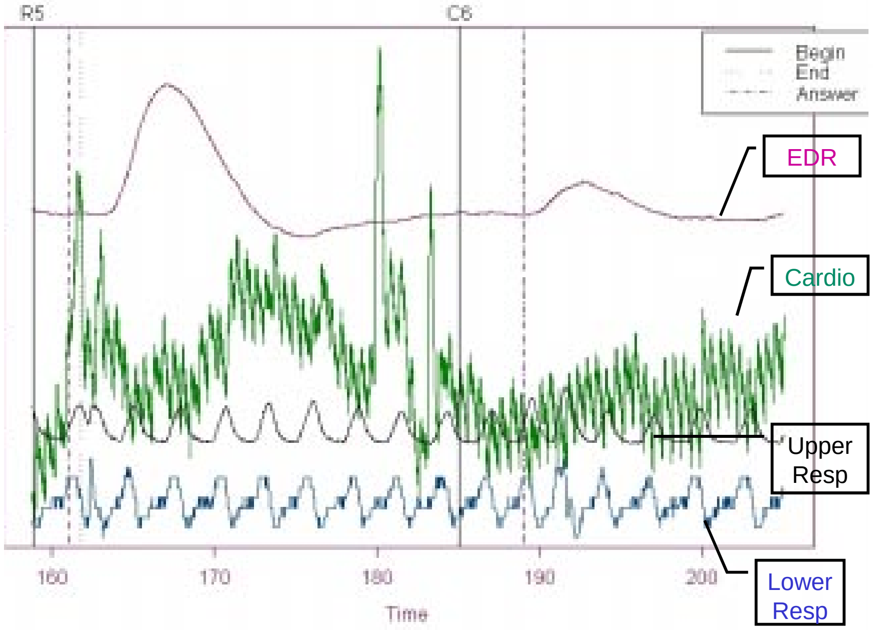


Figure 2: Approximately 40 seconds of data from a guilty person spanning a control and a relevant question. The responsiveness on these two questions is compared. The vertical solid line marks a question onset, the dashed line is the end of the question and dotted line is the answer time.

Electrodermal (EDR) tracing is the most prominent signal. Based on the limited information about the Axciton instrument, our EDR signal is a hybrid of the skin conductance and skin resistance [10] and little of known psychophysiological research can be applied. In manual scoring an evaluator may look for changes in amplitude and duration of response. When there is no reactivity the tracing is almost a horizontal line. Notice in Figure 2 the deceptive person has a higher response after the relevant question than after the control question. Psychophysiology literature reports a 1-3 seconds delay in response to a stimulus. We observe EDR latency of about 1-6 seconds from the question onset.

Research has shown that stronger stimulation elicits larger EDR response, but repetitive stimulation leads to habituation [9]. In Figure 3 notice the decrease in the response on the first relevant question across three charts. This could be a sign of habituation where a response to a stimulus is reduced with repeated exposure to the same question. However, in a number of cases the sequence of the questions may not be the same across the charts so we might be observing different responsiveness to semantically different questions and not habituation. For example, the first relevant question on chart 1 may appear in the third postion on chart 3. In addition, different people have different responsiveness to the same stimulus. It's been found that the skin conductance response (SCR) is correlated with other measures. For example, increase in SCR is associated with increase in blood pressure and heart rate.

Cardiovascular tracing records systolic stroke (pen up), diastolic stroke (pen down) and the dichotic notch. The evaluator looks for changes in baseline, amplitude, rate and changes in dichotic notch (position, disappearance). For cardiovascular activity the blood pressure usually ranges from 80mmHg to 120mmHg [5], but we cannot utilize this knowledge since the scale of our tracings is arbitrary with respect to known physiological measurment units. In fight-or-flight situations, heart rate and blood pressure typically both increase.

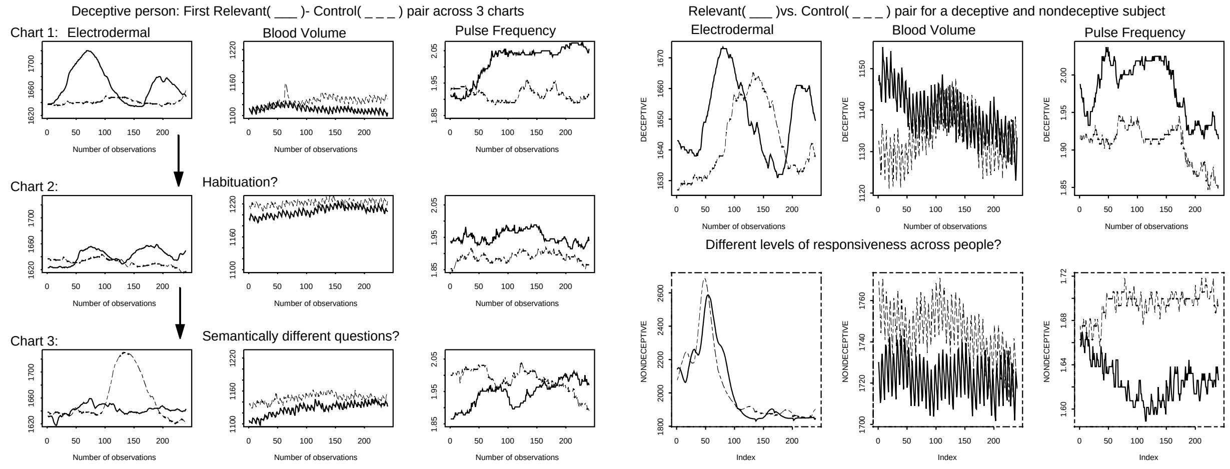


Figure 3: The figure on the left shows overlaid response windows for electrodermal, blood volume and pulse frequency series on charts 1, 2 and 3 of a deceptive person for the first relevant-control question pair. The figure on the right shows the same series on chart 1 of a deceptive and non-deceptive persons for a relevant-control question pair.

Besides habituation there are other issues manifested in these data that may influence feature extraction, evaluation and modeling. Latency differences are present across different question pairs within the same chart and for the same person. In Figure 5, for example, notice how the latency changes for the EDR as we move from the first relevant-control pair to the third. We can also observe different responsiveness (e.g., magnitude) across different questions. This phenomenon, however, may actually be due to a body's tendency to return to homeostasis and not due to a different reaction to different stimuli.

Our analysis revealed diverse test structures even within the same test format. The ZCT usually has the same number of control (C) and relevant (R) questions. A typical sequence is CRCRCR. The MGQT proposed sequence is RRCRRC. These sequences may be interspersed with other question types, and in our data we found at least 15 different sequences. The questions varied greatly across tests and were semantically different among subjects within the same crime. The order of questions varied across charts for the same person. Two problems we faced were the variable number of charts and variable number of relevant questions. Missing relevant questions should be treated as missing data; however, in this project we did not have sufficient evidence to properly impute these values. Thus we chose to drop the fourth relevant-control pair when it existed. For eight subjects who were missing the third relevant-control pair, we replaced their value by zero, i.e., we assumed that there was no difference in the response on that particular relevant-control pair. Elimination of both the fourth chart and the fourth relevant-control pair when missing, and replacement of missing values with zeros did not significantly change the model coefficients nor the final result of classification. These types of differences across cases pose major problems for both within- and between-subject analyses, unless all the responses are averaged.

The crucial information for the development of a statistical classifier of polygraph data is ground truth (i.e., knowledge of whether a subject was truly deceptive or nondeceptive). Ideally, determination of ground truth should be independent of the observed polygraph data, although it is not clear how the ground truth was established for some of our cases. This introduces uncertainty in class labels, in particular for innocent cases since their ground truth is typically set based on somone else's confession. We proceed as though the ground truth in our data is correct.

5 Statistical Analysis

We follow the general framework described in Section 3 for for development and estimation of the logistic regression classification model. The analysis can be broken into Signal Processing, Feature Extraction, Feature Evaluation, Modeling and Classification, and Cross-Validation.

5.1 Signal Processing

With modern digital polygraphs and computerized systems, the analog signals are digitized and the raw digitized electrodermal, cardiovascular and respiratory signals are used in the algorithm development. The primary objective of signal processing is to reduce the noise-to-information ratio. This traditionally involves editing the data (e.g., to detect artifacts and outliers), some signal transformation, and standardization. Our goal is to do a minimal amount of data editing and preserve the raw signal since we lack information on actual instrumentation and any type of filtering performed by either the machine or the examiner.

We first subsampled the 60Hz data by taking every fifth observation for each channel. Next we transformed the cardiovascular recording. We separated the relative blood volume from the pulse, constructing a new series for the relative blood pressure and another one for the pulse frequency. This was done by first calculating the average signal by applying a moving average with a window of size five. This gives a crude measurement of relative blood pressure. The averaged signal was subtracted from the original signal to produce the pulse. The pulse frequency time series is obtained by first removing ultra-high frequency by applying a low pass filter[2]. The filtered signal is made stationary by subtracting its mean. For each window of size 199 observations, we computed the spectral density of a fitted sixth order auto-regressive model[3]. Via linear interpolation we calculated the most prominent frequency. The procedure was repeated for the length of the time series to obtain a new series representing the frequency of a person's pulse during the exam (see Figure 4).

[2] We used the Matlab built-in *Butterworth* filter of the 9th order at frequency 0.8.

[3] We explored different AR models as well, but AR(6) seems to capture the changes sufficiently.

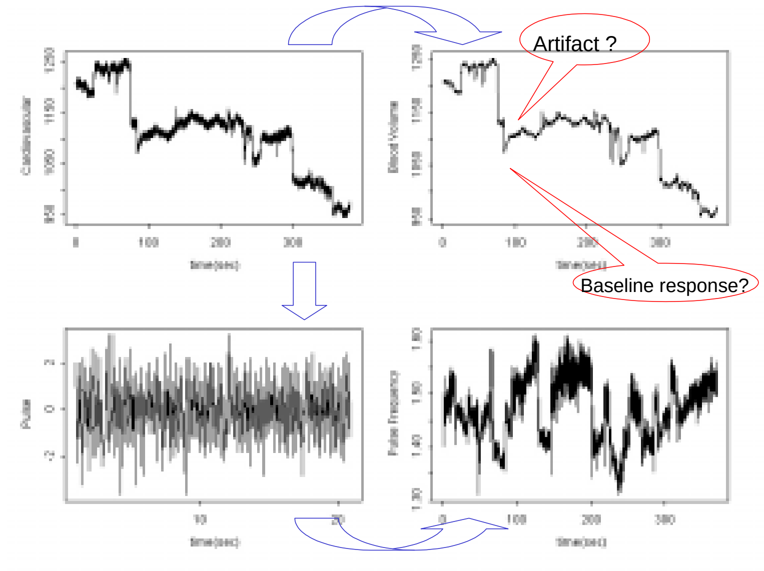


Figure 4: Signal transformation of cardiovascular signal.

5.2 A Simplified Approach to Feature Extraction

The discussion of general statistical methodology for prediction and classification in Section 3 emphasized the importance of feature development and selection. A feature can be anything we measure or compute that represents the emotional signal. Our goal was to reduce the time-series to a small set of features with some relevance in modeling and classifying deception.

Our initial analysis tried to capture low and high frequency changes of the given measurements. To capture slow changes we extracted *integraded differences* and *latency differences* features within a 20-second window from the question onset. Within the same response interval, we extracted *spectral properties differences* to capture high frequency changes. These three features are crude measures of differential activity on relevant and control questions. We used the same features for all signals except for the respiration, where we did not use the latency.

Integrated Differences. The integrated difference is the area between two curves.

$$d_{ijkl} = \sum_{l=1}^{n} R_{ijkl} - C_{ijkl}, \tag{4}$$

is the integrated difference of the i^{th} relevant question versus the i^{th} control question of the j^{th} channel on the k^{th} chart, where $n = 240$ is the number of observations in the response window (see Figure 5).

Latency Differences. We calculated latency for each 20-second window for control and relevant questions on all channels except respiration as follows:

1. Take the absolute value of the differenced time series, $Y_t = |\triangle X_t|$.
2. Calculate the cumulative sum, $Y_j = \sum_{k=0}^{j} X_k$, and normalize it, i.e., $Z_j = \frac{Y_j}{Y_n}$.
3. Define latency as the minimum Z_j such that $Z_j > 0.02$, i.e., $\ell = \min\{Z_j : Z_j \geq 0.02\}$.
4. Define the latency difference feature as the difference in the latency at the relevant and the control questions: $\ell_{rc} = \ell_r - \ell_c$.

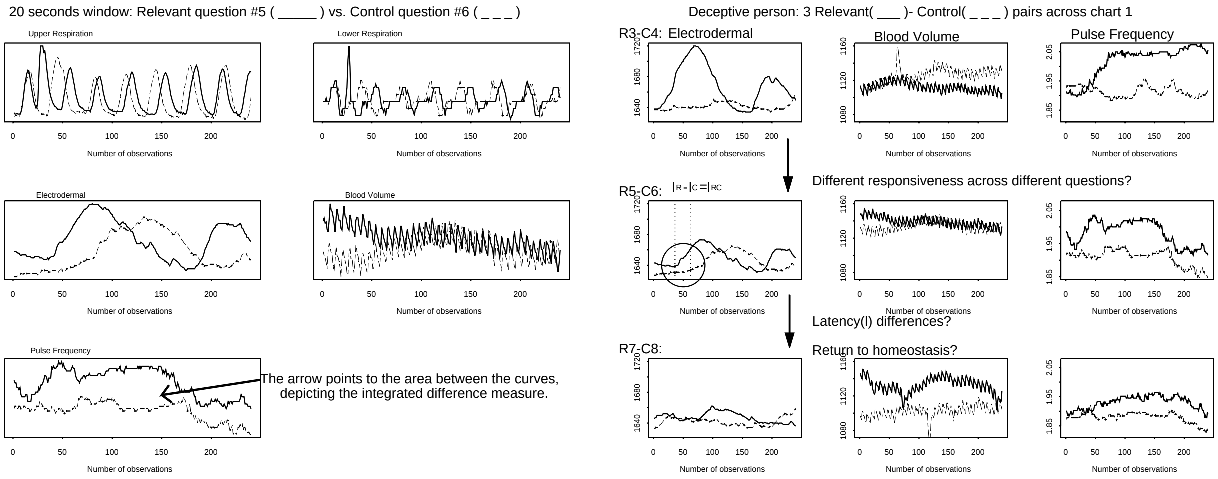


Figure 5: The figure on the left are the overlaid response windows for electrodermal, blood volume, and pulse frequency series on chart 1 of a deceptive person for a relevant-control question pair. The figure on the right shows the same series on chart 1 of a deceptive person for 3 pairs of relevant and control questions.

Spectral Properties Differences. High frequency measure is the difference between spectral properties that we defined in the following way:

1. Apply a high pass filter[4] on a 20-second window for each control and relevant question.
2. Generate periodograms as an estimator measure of spectrum.
3. Assess the spectral properties difference:
   (a) Calculate a mean frequency component, $f_c = \int_0^\pi \lambda S_c(\lambda)\, d\lambda$, where λ is the spectral density of the process, and S_c is the estimated measure of the spectrum.
   (b) Calculate the variance of the frequency component, $v_c = \int_0^\pi \lambda^2 S_c(\lambda)\, d\lambda - f_c^2$.
   (c) Combine (a) and (b) to get $h_{rc} = |f_r - f_c| + \left|\sqrt{v_r} - \sqrt{v_c}\right|$.

These extracted features are measures of responsiveness to the stimuli. For integrated differences and latency differences measures we expect positive values if the response is higher on the relevant question, negative if it's higher on the control questions and zero if there is no difference. Spectral proportion differences only give the magnitude of the differential activity.

5.3 Feature evaluation, modeling, and classification

This section reviews aspects of feature selection and of statistical modeling involving the development of scoring rules into classification rules. The extracted features were considered in three types of comparisons between relevant and control questions:

1. each relevant compared to its nearest control,
2. each relevant compared to the average control,
3. averaged relevant compared to the average control.

[4] We used built-in *Butterworth* filter from Matlab.

In the first two settings the maximum number of continuous variables per subject was 240 (4 relevant-control pairs × 5 channels × 4 charts × 3 features), while the third setting had 60. Since the potential variable space is large relative to the sample size, and since the variables are highly correlated, particularly in the first two settings, we evaluated them graphically, via clustering, principal-component analysis (PCA) and with univariate logistic regression trying to reduce dimnesionality. The remainder of this report will focus on third setting. The other two settings are briefly discussed in Technical Appendix.

Figure 6 shows separation of the two classes given the integrated differences feature for the electrodermal (dEdr) channel or the electrodermal latency differences (lEdr) versus the integrated differences for blood volume (dBv). These are values averaged across charts. Most deceptive subjects (the 1s in the figure) have values greater than zero on dEdr and their distribution is slightly skewed left on pulse frequency (Fqp). Nondeceptive subjects (represented with 0s) mostly have values less than zero on dEdr. They are less variable on Fqp and are centered around zero. Most deceptive subjects have a positive latency differences measure; their latency is longer on relevant than on the control questions (when averaged within and across charts). Nondeceptives show a tendency of having less variable values that are less than zero (i.e., longer latency on EDR on control questions, but not as much between variability as for deceptive subjects). A bivariate plot of the integrated differences for blood volume and pulse frequency shows less clear separation of the two groups. Deceptive subjects show a tendency to have higher blood volume responses on relevant than on control questions while the opposite holds for nondeceptive examinees.

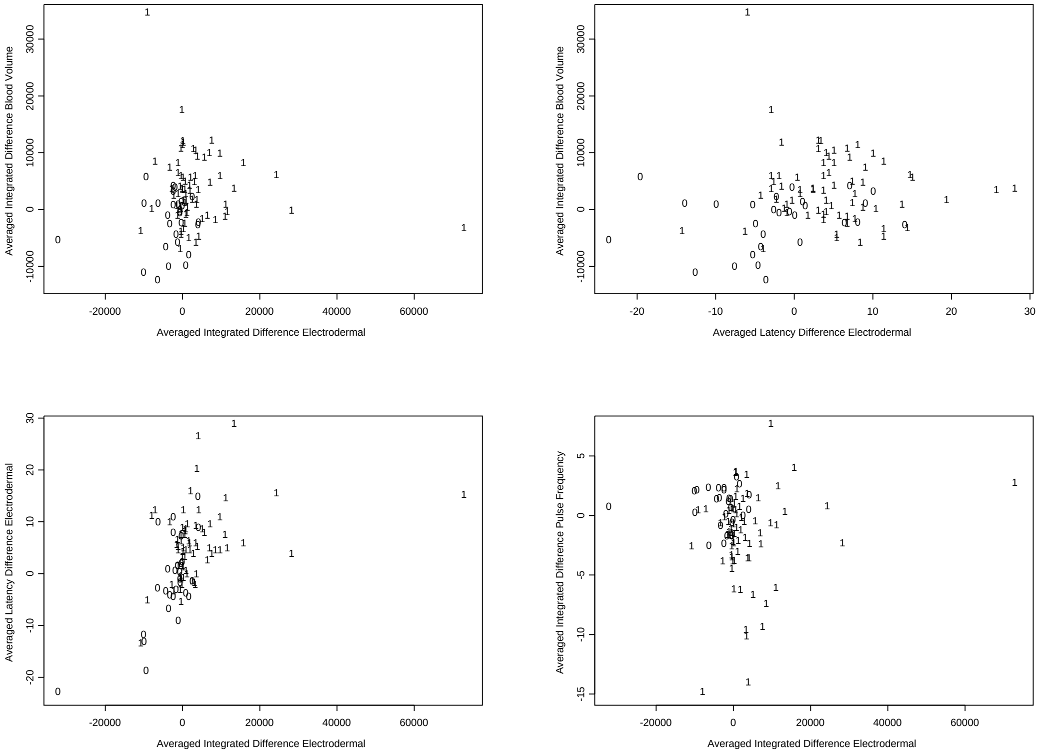


Figure 6: Bivariate plots for some of the averaged features of different channels/signals.

5.4 Logistic Regression

We used data from 97 randomly chosen subjects (69 deceptive and 28 nondeceptive) for evaluation and selection of the best subset of features for the logistic regression classification model. The remaining 52 cases (29 deceptive and 23 nondeceptive) were used for testing. Since the questions across subjects are semantically different and there is no consistent ordering, we developed a model based on comparison of the averaged relevant questions versus the averaged control questions. Logistic regression models developed on variables from the first two settings even when principal components are used as predictors yield multiple models with 8 to 10 predictors. These predictors vary across different models and perform poorly on the test set, although they may achieve perfect separation on the training dataset(see Technical Appendix for brief results).

Average Relevant vs. Average Control. For each chart, each channel and each feature we calculated the average relevant and average control response over the 20-second window. Typically, if we have a noisy signal, one simple solution is to average across trials (in our case across charts) even though we lose some information on measurement variability between different charts.

$\bar{R}_{ij.} = \frac{\sum_{k=1}^{nr_i} R_{ijk}}{nr_i}$ is the averaged relevant response and $\bar{C}_{ij.} = \frac{\sum_{k=1}^{nc_i} C_{ijk}}{nc_i}$ is the averaged control response on the i^{th} chart, j^{th} channel, where nr_i is the number of relevant questions and nc_i is the number of control questions on the i^{th} chart. We calculate the averaged relevant ($\bar{R}_{.j.}$) and control ($\bar{C}_{.j.}$) responses across m charts producing a total of 13 predictors: 5 for integrated differences, 5 for spectral proportion differences and 3 for latency differences.

The logistic regression was performed for each feature independently on each chart, across the charts and then in combination to evaluate the statistical significance of the features. A stepwise procedure in Splus software was used to find the optimal set of features. Neither clustering nor PCA improved the results. The following models are representative of performed analyses on each feature and when combined: Integrated Differences (M1), Latency Differences (M2), Spectral Properties Differences (M3), and All 3 features (M4).

Model	M1	M2	M3	M4
Features	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)
Intercept×10	+4.90 (3.03)	+6.80 (2.50)	+5.46(3.18)	−3.07 (2.96)
Integrated Diff. Electrodermal×10^4	+3.15 (1.02)			+1.59 (0.62)
Integrated Diff. Blood Volume×10^4	+2.14 (0.75)			+1.07 (0.44)
Integrated Diff. Pulse Frequency×10	−3.72 (1.44)			−2.49 (0.87)
Latency Diff. Electrodermal×10		+1.43 (0.404)		+0.35 (0.38)
Spectral Diff. Blood Volume×10^2			+6.78(3.48)	+3.64 (2.32)
Spectral Diff. Respiration×10			+1.43(2.33)	
Spectral Diff. Pulse Frequency			−261.48(364.48)	

Table 2: Features with the estimated logistic regression coefficients and standard errors for models M1, M2 and M4. A positive sign for a weight indicates an increase in deception, while a negative sign denotes decrease. The absolute value of a weight suggests something about the strenght of the linear association with deception.

We considered models on individual charts and observed almost identical models across charts. Chart 3 did worse on cross-validation than the other two charts, and relied more on high frequency measures of respiration and sweating. Chart 2 added to the detection of innocent subjects in comparison to chart 1. For chart 2 the latency differences on blood volume was a slightly better predictor than the high frequency measure which is more significant on chart 1. Table A7 in the

Appendix gives estimated coefficients and their standard erros for the models that following models that performed the best on each chart:

$$\text{Chart1 - M5: Score} = \hat{\beta}_0 + \hat{\beta}_1\text{dEdr} + \hat{\beta}_2\text{dBv} + \hat{\beta}_3\text{dFqp} + \hat{\beta}_4\text{hPn2} + \hat{\beta}_5\text{hBv} + \hat{\beta}_6\text{lEdr}, \tag{5}$$

$$\text{Chart2 - M6: Score} = \hat{\beta}_0 + \hat{\beta}_1\text{dEdr} + \hat{\beta}_2\text{dBv} + \hat{\beta}_3\text{dFqp} + \hat{\beta}_4\text{lEdr} + \hat{\beta}_5\text{lBv}, \tag{6}$$

$$\text{Chart3 - M7: Score} = \hat{\beta}_0 + \hat{\beta}_1\text{dEdr} + \hat{\beta}_2\text{dBv} + \hat{\beta}_3\text{hPn2} + \hat{\beta}_4\text{hEdr} + \hat{\beta}_5\text{lEdr}. \tag{7}$$

The linear combination of integrated differences was the strongest discriminator. Latency had the most power on the electrodermal response. Our high frequency feature on any of the measurements was a poor individual predictor, particularly on nondeceptive people, however it seems to have some effect when combined with the other two features. All features show better discrimination on electrodermal response, blood volume and pulse frequency than on the respiration measurements.

5.5 Classification Results

We tested the previously described models for their predictive power on an independent test set of 52 subjects. This is known as hold-out-set cross validation. Table 3 summarizes the classification results based on a 0.5 probability cutoff. A probability of 0.5 or above indicates deception, and a probability less than 0.5 indicates truthfulness.

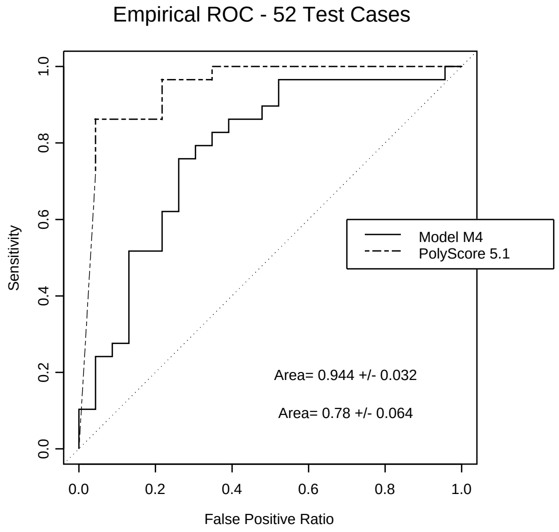
Model	*Training*		*Test*	
	Deceptive(%)	Nondeceptive(%)	Deceptive(%)	Nondeceptive(%)
M1	94	64	97	52
M2	96	29	90	9
M3	99	7	100	4
M4	93	61	97	48
M5	92	50	83	38
M6	88	43	87	52
M7	92	57	83	22

Table 3: Percents of correctly classified subjects from hold-out-set cross validation.

We ran the same subsets of training and test data through the Polyscore. Figure 6 shows receiver operating characteristic curves (ROCs) of model M4 performance and of PolyScore 5.1 on 52 test cases. This is not an independent evaluation of PolyScore algorithm since some of these cases were used in its development. ROC and area under the curve give quantitative assessments of a classifier's degree of accuracy. It was shown by [8] that ROC overestimates the performance of the logistic regression classifier when the same data are used to fit the score and to calculate the ROC.

	Training (N=119)				Test(N=30)			
	Deceptive%		Nondeceptive%		Deceptive%		Nondeceptive%	
	M1	M4	M1	M4	M1	M4	M1	M4
Mean	91	90	60	57	92	92	70	66
St.Error	4.1	2.1	2.3	2.1	5.4	5.6	21.6	21.5

Table 4: Percent of correctly classified individuals in k-fold cross validation at 0.5 cutoff value.



≥ 0.5 cutoff	M4	PolyScore 5.1	Others*
Deceptive	90%	86%	73-100%
NonDeceptive	48%	78%	53-90%

Figure 7: ROCs for classification results of PolyScore 5.1 and M4 on 52 test cases.The table shows percent correct when 0.5 is considerd as a cutoff value. In practice PolyScore 5.1 uses 0.95 and 0.05 as the cutoff values for classifiying deceptive and nondeceptive subjects (everything inbetween is considered inconclusive). (*)These values are based on different cutoff values and range over the percent correct when inconclusives are included and excluded giving higher percent correct when incoclusive cases are excluded.

Since k-fold cross-validation works better for small data sets [17] we performed 5-fold cross-validation on models M1 and M4. The results are presented in Table 4. The number of innocent subjects in training runs from 30 to 47 out of 51, and deceptive from 72 to 80 out of 98 (see Table A6). We belive that 100% correct classification in the first run, which is highly inconsistent with the other runs, was due to the small number of nondeceptive test cases in that run. The average area under the ROC for M4 is 0.899(±0.05). When we apply shrinkage correction proposed by [8] the average area under the curve is approximately 0.851.

6 Discussion

The objective of the automated scoring algorithms for polygraph data is to create reliable and statistically valid classification schemes minimizing both false positive and false negative rates. Beginning in the 1970s, various papers in the polygraph literature offered evidence claiming to show that automated classification methods and algorithms for analyzing polygraph charts could do so. According to [10] the accuracies of five different computer algorithms range from 73% to 89% on deceptive subjects when inconclusives are included and 91% to 98% when they are excluded. For innocent subjects these numbers vary from 53% to 68%, and 72% to 90%.

Our analyses based on a set of 149 criminal cases provided by DoDPI suggests that it is easy to develop such algorithms with comparable recognition rates. Our results are based on a 0.5 probability cutoff for two groups: deceptive (probability greater than 0.5) and nondeceptive (probability less than 0.5). Other cutoff values would allow us to balance the errors differently. Neither clustering nor PCA significantly improve our results, which is consistent with the recent work of [6].

One possible explanation for the relatively poor classification preformance is the small sample size and, in particular, the small number of nondeceptive cases. However, PolyScore algorithms, for example, had a much larger database [25] but thier accuracy rates are not significantly better than ours. Since a stepwise procedure for selecting the variables relies predominantly on the random variability of the training data, we expect to get models more specific to the training sample which do not necessarily generalize well to the test sample.

Another possible explanation could be high variability and presence of measurment errors that come with real-life polygraph data, where there is a lack of standards in data collection and recording. Our exploratory data analysis points to problems with question inconsistency, response variability within an individual and across individuals (due to nature, gender, etc.), and possible learning effects. It is not always clear where differences in responses come from; are we dealing with habituation or comparing semantically different questions across the charts and hence having different responsivness to different questions? Since in our data questions are semantically different, and no consistent ordering within and across charts could be established, we averaged the relevant and control responses and then look at their difference. PolyScore and CPS algorithms take the same approach. This methodology ignores the question semantics which could be a flaw in this approach. These phenomena could be better studied in the screening type tests or with more standardized laboratory cases[5] where there is consistency in the order and the type of questions asked within and across both charts and individuals. Although CPS algorithms have been developed on laboratory data, they have not achieved significantly better results.

Each step of the analysis outlined in Section 5 has numerous issues worth further exploration and deliberation. Below we discuss a selected few.

Signal processing. In the signal processing stage we extract blood volume and pulse from the cardiovascular recording. While other algorithms have performed similar transformations, we are not aware that they produce pulse frequency and use it as an additional measurement which in our analyses significantly aids in classification. Cardiovascular signal could be further purified by removing a frequency band due to respiration from 0.12Hz to 0.4Hz. At this time we are not sure if this would improve the results.

Dawson et al. [9] point out the large variability across individuals in their EDR responsiveness. They recommend standardizing EDR data. The recommendations, however, differ for skin conductance response and skin conductance level (SCL). Some of the elements needed for the proposed standardization are not easily identified in our data. The minimum SCL should be calculated from the rest period and the maximum during some stronger activity. What could we consider a rest period: an irrelevant question for either deceptive or non-deceptive? Which irrelevant would it be: the first or any other? The two algorithms address different levels of responsiveness across different people by standardizing the features. The standardization is slightly problematic because of all these issues. Second, the standardization is typically done within the response window only and not on the complete waveform.

Psychophysiological literature suggests that artifacts, outliers and other signal noise should be removed. However, it is not always clear if something is an artifact or actual reaction to a stimulus as depicted in Figure 4. In this work we did not perform artifact detection and removal. More on this important topic related to automated scoring of polygraph data may be found in [22, 13, 7].

Feature extraction. An important part of feature extraction is defining an event (response

[5] See discussion on possible downfalls of lab data in [7].

interval) that captures information on deception. It is not trivial to determine where the reaction begins and ends with respect to the signal and the feature. The analyses presented in this report rely on a 20-second window from the onset of the question. A fixed window size may not account for the fact that different signals by their nature have different frequencies and a person is likely to show faster or slower reactions on one channel than the others. Windows of different sizes for different channels need to be considered. Although JHUAPL uses different window sizes, there is a lack of scientifically recorded procedures that they used for coming up with these windows. We consider the window only from the question onset. Psychophysiological research, for example for the EDR, indicates that the reactions should be considered by looking a few seconds before the onset of the question. Another issue that appears to be ignored by the current algorithms is the possibility of anticipation and learning that may have occurred during the examination. If the order of the questions remains the same across the three repetitions, the subject is likely, at least by the third chart, to anticipate the upcoming question and react sooner than expected. We observed some cases where the answers occur before the asking of the question ends. This could be a false phenomena, and simply an error of the examiner in recording the time of the answer.

Feature evaluation and selection. Discussion of general statistical methodology for prediction and classification at the beginning of this report emphasized the importance of feature development and selection. With three relatively crude and simple features we tried to capture both low frequency (integrated difference features, and latency difference features) and high frequency (spectral properties difference) changes in the physiological recordings during the exam. The general psychophysiological literature suggests a larger array of possible features such as skin conductance level or half-recovery time [9]. Since we have some hybrid of skin conductance and skin resistance, it is questionable if and how any of suggested features apply on our data. Similarly, cardiovascular activity is typically analyzed using heart rate and its derivatives such as the heart rate variability or the difference of the maximum and minimum amplitudes. Brownley et al.[5], however, state that reliability of heart rate variability as a measure is controversial and they suggest the use of respiratory sinus arrhythmia (RSA), which represents the covariance between the respiratory and heart rate activity. This implies a need for the frequency-domain analysis in addition to the time-domain analysis of the biological signals. Harver et al. [16] suggest looking at the respiratory rate and breathing amplitude as possible features describing respiratory responses. They also point out that recording changes only of upper or lower respirations is not adequate to estimate relative breathing amplitude.

In general, area measures (integrated activity over time) are less susceptible to high-frequency noise than the peak measures, but the amplitude measurements are more reliable than the latency [12]. Early research focusing specifically on the detection of deception suggested that the area under the curve and amplitudes of both skin conductance and cardiovascular response can discriminate between the deceptive and truthful subjects. Other features investigated included duration of rise to peak amplitude, recovery of the baseline and the overall duration of the response. [21] report that line-length, the sum of absolute differences between adjacent sample points, which captures some combination of rate and amplitude is a good measure of respiration suppression. Our analyses indicate that integrated differences have strong classification power on all of the channels but the least on the respiration. Similar inference can be made for the latency difference measure, which is less significant than the area measure.

Computerized analysis of digitized signal offers a much larger pool of features, some of them not easily observable by visual inspection, but large pool of features raises problems in feature selection. Polyscore for example has considered on the order of 10,000 possible features. The statistical and data-mining literatures are rife with descriptions of stepwise and other feature selection procedures, but the multiplicity of models to be considered grows as one considers transformations of features and interactions among features. All of these aspects are intertwined and thus the methodological literature fails to provide a simple and unique way to achieve the empirical objectives of identifying a subset of features in the context of a specific scoring model that has good behavior when used on a new data set. When the number of features is larger, the exhaustive approach is clearly not feasible. If one has a small training set of test data (repeatedly uses the same test data) one may obtain features that are well suited for that particular training or test data, but that are not the best features set in general. What most statisticians argue is that fewer relevant variables do better on cross-validation, but even this claim comes under challenge by those who argue for model-free, black-box approaches to prediction models (e.g., see [4]). In our analyses the most consistent model has only 3 variables and it is easily interpretable. PolyScore 3.2 ultimately ended up using 10 features for logistic regression: 3 describing GSR, 3 for blood volume, 2 for pulse and 2 describing respiration but the authors do not disclose precisely which features these are [25]. PolyScore version 5.1, based on a neural network, purports to use 22 features. The most recent version of the CPS algorithm, uses only 3 features: SC amplitude, the amplitude of increases in the baseline of the cardiograph and a line-length composite measure of thoracic and abdominal respiration excursion [21]. In the present setting, the number of cases used to develop and test models for the algorithms under review was sufficiently small that the seeming advantages of these data-mining approaches are difficult to realize.

Classification. Logistic regression applies maximum likelihood estimation to the logit transform of the dependent variable. Since it estimates log odds of a person being deceptive, it actually gives us a probably of deception. Logistic regression is relatively robust since it doesn't assume linearity between covariates and the response, covariates do not need to be normally distributed, and it doesn't assume homoscedasticity of the response, and does not need normally distributed errors. Currently we assign equal weights to all of our covariates. It is possible that the different features and or signals should have more weight. One main difference between manual and automatic scoring is that the manual scoring equally weights all three channels, while the automated scoring more heavily weights the electrodermal response. JHUAPL algorithms assign different weights to features associated with different signals. From anecdotal experience it has been noted that the electrodermal (EDR) output is more easily scored than the other channels and that examiners heavily rely on EDR signal. It also appears that respiratory recording bears the least weight, although some recent results point otherwise [6]. In our analyses, we also observed that features on EDR have the most classification power whereas the ones on respiration have the least.

The models, as expected, generalize poorly to the test set. We observe the increase in lack of fit as we move from the averaged models to pairwise model. This might be due to the fact that we do not account for the ordering of the questions; neither manual or other automatic scorings do. It also appears that we need more predictors to describe the deceptive and non-deceptive classes as we move from the averaged to pairwise models. When based on a large number of classifying variables we appear to achieve perfect separation of deceptive and nondeceptive individuals on training data but perform poorly on the test set. Statisticians have recognized the problem with such "overfitting" of the data, and shown that the performance of these classifiers often deteriorates badly under proper cross-validation assessment (see [17], for a general discussion of feature selection and the discussion that follows for specifics in the polygraph setting). These "complex" algorithms often turn out to be less effective on a new set of cases than those based on a small set of simple features.

All the above discussed issues point to difficulty of capturing all the variability and the lack of structure in the specific incident polygraphs, and in producing a generalizable model. Perhaps it is not reasonable to expect that a single algorithm will successfully be able to detect guilty and innocent examinees. The solution may lay in the data collection and on detailed research on underlying theory for polygraphs, before proper statistical modeling can be effectively utilized. On the other hand polygraph testing in a screening setting is more structured for an individual and across individuals. TES asks the same questions on two charts in the same order for all subjects. From our analyses it could be reasonable to expect that a generalizable model can be developed for screening purposes. Of course having a generalizable model does not mean that we will have either validity or accuracy. However, the issue here is if we are going to have strong enough differential responsiveness since the questions are less specific.

Finally, in the cases we examined there is little or no information available to control for selection bias, assesment of ground truth, differences among examiners, examiner-examinee interactions, and delays in the timing of questions. Most of these are not addressed by current scoring algorithms. More discussion on these issues and inflated accuracies can be found in Appendix F of [7]. Further if we are to extend this results to screening data, we need to be cautious since we might be dealing with two completely different populations (i.e. common thief, vs. trained secret agents). Some of these problems can be overcome by careful systematic collection of polygraph field data, especially in a screening setting, and others cannot. Controlling for all possible dimensions of variation in a computer-scoring algorithm, however, is a daunting task unless one has a large database of cases.

7 Conclusion

This report presents an initial evaluation and analysis of polygraph data for a set of real-life specific incident cases. With a very simple approach we have managed to obtain accuracy rates comparable to what's currently being reported by other algorithms and manual scoring. The fact that we are able to produce a number of different models that account for different predictors yet give similar results, points to the complexity that underlines assessment and/or classification of examinee's deceptiveness.

This work can be redefined and extended in a number of ways. More features could be extracted and explored. Thus far these efforts have not resulted in significantly smaller errors, hence it raises a question how far could this approach go beside what's already has been reported. One could imagine improvements to current methods by running a proper Bayesian analysis and incorporating prior knowledge on prevalence. Our inclination would be to do a more complex time series analysis of these data. The waveform of each channel can be considered and the analysis would gear towards describing a physiological signature for deceptive and nondeceptive classes. Clearly the ordering of the questions should be accounted for. A mixed-effects model with repeated measures would be another approach, where repetitions would be measurements across different charts. In other areas with similar data researchers have explored the use of Hidden Markov Models [11].

There has yet to be a proper independent evaluation of computer scoring algorithms on a suitably selected set of cases, for either specific incidents or security screening, which would allow one to accurately assess the validity and accuracy of these algorithms. One could argue that computerized algorithms should be able to analyze the data better because they use the tasks which are more difficult even for a trained examiner to perform, including filtering, transformation, calculating signal derivatives, manipulating signals, and looking at the bigger pictures not merely adjacent comparisons. Moreover, computer systems never get careless or tired. However, success of both numerical and computerized systems still depends heavily on the pre-test phase of the examination. How well examiners formulate the questions inevitably affects the quality of information recorded. We believe that substantial improvements to current numerical scoring may be possible, but the ultimate potential of computerized scoring systems depends on the quality of the data available for system development and application, and the uniformity of the examination formats with which the systems are designed to deal.

Acknowledgments

The author would like to thank Stephen E. Fienberg and Anthony Brockwell, for their advice and support on this project. Special thank you to the members of the NAS/NRC Committee to review the scientific evidence on polygraph for the opportunity to work with them, and to Andrew Ryan and Andrew Dollins of Department of Defense Polygraph Institute for providing the data.

References

[1] Alder, K. 1998. To Tell the Truth: The Polygraph Exam and the Marketing of American Expertise. *Historical Reflections.* Vol. 24., No. 3, pp.487-525.

[2] American Polygraph Association. *www.polygraph.org*

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

[4] Breiman, L. 2001. Statistical modeling: The two cultures (with discussion). *Statistical Science* 16:199-231.

[5] Brownley, K.A., B.E. Hurwitz, N. Schneiderman. 2000. Cardiovascular psychophysiology. Ch. 9, pp. 224-264, in *Handbook of Psychophysiology*, 2nd Ed., J.T. Cacioppo, L.G. Tassinary, and G.G. Bernston, eds. NY: Cambridge University Press.

[6] Campbell, J.L. 2001. *Individual Differences in Patterns of Physiological Activation and Their Effects on Computer Diagnoses of Truth and Deception. Doctoral Disseration.* The University of Utah.

[7] Committee to Review the Scientific Evidence on the Polygraph. 2002. *The Polygraph and Lie Detection.* National Academy Press, Washington, DC.

[8] Copas, J.B., Corbett, P. 2002. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*, 89(2), pp. 315-331.

[9] Dawson, M., A.M. Schell, D.L. Filion. 2000. The electrodermal system. Ch. 8, pp. 200-223, in *Handbook of Psychophysiology*, 2nd Ed., J.T. Cacioppo, L.G. Tassinary, and G.G. Bernston, eds. NY: Cambridge University Press.

[10] Dollins, A.B., D.J. Kraphol, D.W. Dutton. 2000. Computer Algorithm Comparison. *Polygraph*, 29(3).

[11] Fernandez, Raul. 1997. *Stochastic Modeling of Physiological Signals with Hidden Markov Models: A Step Toward Frustration Detection in Human-Computer Interfaces.* Master's Thesis. MIT.

[12] Gratton, G. 2000. Biosignal Processing. Ch. 33, pp. 900-923, in *Handbook of Psychophysiology*, 2nd Ed., J.T. Cacioppo, L.G. Tassinary, and G.G. Bernston, eds. NY: Cambridge University Press.

[13] Harris, J.C., Olsen, D.E. 1994. Polygraph Automated Scoring System. U.S. Patent #5,327,899.

[14] Harris, J. 1996. Real Crime Validation of the PolyScore(r) 3.0 Zone Comparison Scoring Algorithm. JHUAPL.

[15] Harris, J. 2001. Visit by Aleksandra Slavkovic, consultant to the Committee to Review the Scientific Evidence on the Polygraph, to The Johns Hopkins University Applied Physics Laboratory, June 18, 2001.

[16] Harver, A., T.S. Lorig. 2000. Respiration. Ch. 10, pp. 265-293, in *Handbook of Psychophysiology*, 2nd Ed., J.T. Cacioppo, L.G. Tassinary, and G.G. Bernston, eds. NY: Cambridge University Press.

[17] Hastie, T., R. Tibshirani, J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* NY: Springer-Verlag.

[18] Hosmer, D.W, Lemeshow, Jr. S. 1989. *Applied Logistic Regression.* John Wiley & Sons, NY.

[19] Jennings, R.J., L.A. 2000. Salient method, design, and analysis concerns. Ch. 32, pp. 870-899, in *Handbook of Psychophysiology*, 2nd Ed., J.T. Cacioppo, L.G. Tassinary, and G.G. Bernston, eds. NY: Cambridge University Press.

[20] Johnson, R.A., and Wichern,D.W. 1992. *Applied Multivariate Statistical Analysis.* Third Edition. Englewood Cliffs, NJ: Prentice-Hall, Inc.

[21] Kircher, J.C., and D.C. Raskin. 1988. Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology* 73:291-302.

[22] Kircher, J.C., and D.C. Raskin. 2002. Computer methods for the psychophysiological detection of deception. Chapter 11, pp. 287-326, in *Handbook of Polygraph Testing*, M. Kleiner, ed. London: Academic Press.

[23] Matte, J.A. 1996. *Forensic Psychophysiology Using Polygraph-Scientific Truth Verification Lie Detection.* Williamsville, NY: J.A.M. Publications.

[24] McLachlan, G.J. 1992. *Discriminant Analysis and Statistical Pattern Recognition.* New York: John Wiley & Sons, Inc.

[25] Olsen, D.E, Harris, J.C., Capps, M.H., Ansley, N. 1997. Computerized Polygraph Scoring System. *Journal Forensic Science* 42(1):61-71.

[26] Raskin, D.C., Hont, C.R., Kircher, J.C. 1997. The scientific status of research on polygraph techniques: The case for polygraph tests. In D.L. Faigman, D.Kaye, M.J. Saks, J. Senders (Eds.) *Modern sceintific evidence: The law and science of expert evidence.* St.Paul, MN:West.

[27] Porges, S.W., R.A. Johnson, J.C. Kircher, and R.A. Stern. 1996. Unpublished Report of Peer Review of Johns Hopkins University/Applied Physics Laboratory to the Central Intelligence Agency.

[28] Swinford,J. 1999. Manually Scoring Polygraph Charts Utilizing the Seven-Position Numerical Analysis Scale at the Department of Defense Polygraph Institute. *Polygraph*, 28(1). pp. 10-27.

8 Technical Appendix

	Example questions of ZCT format
X	This test is about to begin.
1	Is today Wednesday?
2	Regarding that stolen property, do you intend to answer truthfully each question about that?
3	Are you convinced that I will not ask you a surprise question on this test?
4C	Before 1995, did you ever steal anything from an employer?
5R	Did you steal any of that missing property?
6C	Before 1995, did you ever steal anything?
7R	Did you steal any of that missing property from building —— ?
8	Is there something else you are afraid I will ask you a question about, even though I told you I would not?
9C	Before 1995, did you ever steal something and not get caught?
10R	Do you know for sure who stole any of that missing property?
XX	Test is over...

	Example questions of MGQT format
X	This test is about to begin...
1	Is your last name ———?
2	Is your first name ————?
3R	Did you help anyone steal any of that missing property?
4	Are you sitting down?
5R	Did you steal any of those ———?
6C	Other than what you told me, before 1995, did you ever steal anything?
7	Are the light on in this room?
8R	Did you steal any of those ———?
9R	Do you know for sure who stole any of that missing equipment?
10C	Other than what you told me, before 1995, did you ever steal anything and not get caught?
XX	This test is over.

Table A1. Example questions of ZCT and MGQT formats.

	Example of TES questions on Chart 1 and Chart 2:
11	Are you now sitting down?
12	Are the lights on in this room?
SR	Do you intend to answer the security questions truthfully?
1c1	Did you ever violate a traffic law?
1r1	Have you committed sabotage against the United States?
1r2	Have you ever been involved in espionage against the united states?
1c2	Did you ever lose your temper?
2r1	Have you committed sabotage against the united states?
2r2	Have you ever been involved in espionage against the united states?
2c1	Did you ever violate a traffic law?
3r1	Have you committed sabotage against the united states?
3r2	Have you ever been involved in espionage against the united states?
2c2	Did you ever lose your temper?
1I1	Are you now sitting down?
1I2	Are the lights on in this room?
SR	Do you intend to answer the security questions truthfully?
1c1	Did you ever lose your temper?
1r3	Have you illegally disclosed classified information to unauthorized person?
1r4	Have you had any unauthorized foreign contacts?
1c2	Did you ever violate a traffic law?
2r3	Have you illegally disclosed classified information to unauthorized person?
2r4	Have you had any unauthorized foreign contacts?
2c1	Did you ever lose your tamper?
3r3	Have you committed sabotage against the united states?
3r4	Have you ever been involved in espionage against the united states?
2c2	Did you ever violate a traffic law?

Table A2. Example of TES question sequence. The top part are questions asked on the first chart and the bottom are the questions asked on the second chart.

Algorithm	*Deceptive (n=56)*			*NonDeceptive(n=41)*		
	Correct	Incorrect	Inconclusives	Correct	Incorrect	Inconclusives
CPS	41	4	11	28	3	10
PolyScore	49	1	6	26	7	8
Axcon	50	1	5	24	9	8
Chart Analysis	49	2	5	22	8	11
Identifi	49	1	6	22	8	11

Table A3. Number of correct, incorrect, and inconclusive decisions by subject's ground truth for five evaluated alogorithms (see [10])

RRCRRC	RRCRCR	RRCRRCC	RCRCRCR	RCRCRCRC
CRCRC	CRCRRC	CRCRCR	CCRCRC	CRCRCRC
CRCRCCC	CRCRRCR	CRCRCRCC	CRCRCRCR	CRCRCRCRC

Table A4. The possible sequences of relevant (R) and control (C) questions on one chart. When we account for irrelevant questions the number of possible combinations increases.

Features(PolyScore)	Weights($\hat{\beta}$)		Features(CPS)	Weights($\hat{\beta}$)
GSR Range	+5.51		SC Amplitude	+.77
Blood Volume Derivative 75th Percentile	+3.06		SC full recovery time	+.27
Upper Respiration 80th Percentile	-2.60		EBF	+.28
Pulse Line Length	-2.087		BP Amplitude	+.22
Pulse 55th Percentile	+2.16		Respiration Length	-.40

Table A5. List of implemented features and their estimated coefficients in version 3.0 of PolyScore [25], and in CPS [21]. Note that these are on different scale.

Run	*Training*		*Test*	
	Deceptive	Nondeceptive	Deceptive	Nondeceptive
1	72	47	26	4
2	80	39	18	12
3	80	39	18	12
4	79	40	19	11
5	76	43	22	30

Table A6. Training and test sample sizes for 5 runs in cross-validation by ground truth. Each run has 119 training and 30 test cases, 98 deceptive and 51 innocent.

8.1 Missing Data and Clean up

We put a considerable effort in data clean up to allow further automatic processing and statistical analyses. Tcl/Tk scripts were written to extract the data fields from the provided text files. The main problems we encontered were inconsistent labeling, variable test formats and missing relevant information, e.g. there were various representations of the first irrelevant question: I1, IR1, 1, 1I, 1IR, etc. Similar problem existed with other questions.

8.2 Average Relevant vs. Average Control

Model	Chart1(M5)	Chart2	Chart3
Features	Weights $\hat{\beta}$ (SE)	Weights $\hat{\beta}$ (SE)	Weights $\hat{\beta}$ (SE)
Intercept$\times$10	+5.82 (3.56)	+7.89 (2.83)	+6.21(3.19)
Integrated Difference Electrodermal$\times 10^5$	+4.73 (3.11)	+18.1 (4.17)	+18.8(8.54)
Integrated Difference Blood Volume$\times 10^4$	+1.25 (0.47)	+0.80 (0.34)	+2.14(0.72)
Integrated Difference Pulse Frequency$\times$10	−2.12 (0.75)	−1.82 (0.65)	
Latency Difference Electrodermal$\times 10^2$	+4.59 (3.14)	+4.73 (3.31)	+8.57(3.49)
Latency Differece Blood Volume$\times$10		+1.72 (1.04)	
Spectral Difference Blood Volume$\times 10^2$	+5.86 (3.43)		
Spectral Difference Respiration$\times$10	−4.12 (2.56)		+2.76(1.63)
Spectral Difference Electrodermal$\times$10			−11.15(7.95)

Table A7. List of features with their estimated coefficients and standard errors for models based on each chart.

The best model for chart 2 has four of the same features as on chart 1, but instead of high frequency measure on blood volume it has latency measure for blood volume. The classification rates drop slightly for deceptive subjects in comparison to model M5, but the model does better on nondeceptive subjects. It seems that chart 2 data adds some information in particular for nondeceptive subjects to chart 1, but we could use just chart 1 do get similar results as when we combine the data from all three charts.

Models based on data from chart 3 do not do as well as models based on chart 1. They do better than chart 2 on classifying deceptive and worse on classifying nondeceptive subjects. These models also seem to rely more on high frequency measure for respiration and sweating than data from charts 1 and chart 2.

8.3 Each Relevant vs. Average Control

Difficulty in this setup is to reduce the number of potential predictors to obtain a model with the small number of features. The model below is one of the better models when averaged across the charts that still relies on nine predictors. Sensitivity on training and test data are 88.4% and 86.2% while specificities are 67.9% and 47.8%. Latencies are correlated, for example, on Edr between questions one, two and three.

Freatures	Weights $\hat{\beta}$ (SE)
Intercept$\times$10	+5.35 (4.30)
dEdr1 $\times 10^4$	+1.18(0.530)
dEdr2 $\times 10^4$	+1.57(0.75)
dBv2 $\times 10^4$	+1.06(0.49)
dFqp1 $\times$10	−1.12(0.782)
dFqp2 $\times$10	−1.71(1.09)
hPn11 $\times 10^2$	−5.61(10.01)
hPn13 $\times$10	+1.16(1.44)
lEdr3 $\times$10	+1.08(0.63)
lFqp1 $\times$10	−2.88(1.82)

Table A8. Estimated coefficients and their standard errors of a model based on each relevant to averaged control comparison. 'dEdr' - integrated differences feature, 'dBv'- integrated differences blood volume, 'dFqp'- integrated differences pulse frequency, 'hPn1' - spectral properties differences upper respiration, 'lEdr' - latency differences electrodermal, 'lFqp' - latency difference pulse frequency. A digit at the end of each

predictor denotes a relevant question for which the feature was calculated (e.g. 'dEdr1'- integrated differences measure on electrodermal signal for the first relevant question when compared to the averaged control).

When models are considered on each chart individually we notice that on Chart 1 most significant features are on relevant question 2. On Chart 2 significant features are on questions 1 and 3. Chart 2 seems to do a slightly better than chart 1 for integrated differences. On chart 3 it's harder to say if there are any dominating questions. We only considered differences in questions by their postions, and not thier semantics.

8.4 Each Relevant vs. Nearest Control

Various models are considered here. The best classifications rates are obtained when we calculate the relevant-control difference and then average across the charts. We considered models based only on question position. In the preliminary analysis with 97 training cases model relying only on the second question does slightly better than either model based on the first or third relevant question, however, the differences are not significant.

Model below is an example of many similar models that perform 100% correct classification on the training data, but perform significantly worse on the test data. These models are based on features when we do not average across the charts. This particular model is only based on integrated difference feature. 'rc1Edr1' for example stands for integrated difference measure on the first relevant-control question on electrodermal channel on chart 1.

$$Score = \hat{\beta}_0 + \hat{\beta}_1 rc1Edr1 + \hat{\beta}_2 rc1Fqp3 + \hat{\beta}_3 rc2Edr3 + \hat{\beta}_4 rc2Bv2 + \hat{\beta}_5 rc2Fqp1 + \hat{\beta}_6 rc2Fqp2 + \hat{\beta}_7 rc3Pn21 + \hat{\beta}_8 rc3Edr1 + \hat{\beta}_9 rc3Edr2 + \hat{\beta}_{10} rc3Bv1 + \hat{\beta}_{11} rc3Fqp2$$