Predicting Performance Using Galvanic Skin Response

by

Lee Carter Mundell
A.B., University of Chicago (2010)

Submitted to the Sloan School of Management in partial fulfillment of the requirements for the degree of Master of Science in Operations Research at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY June 2016

© Massachusetts Institute of Technology 2016. All rights reserved.

Signature redacted

Signature of Author .......................................................... Sloan School of Management

Signature redacted

Certified by

Signature redacted

May 11, 2016

Certified by ..................

Signature redacted

Tauhid Zaman
Assistant Professor
Thesis Supervisor

Certified by ..................

Signature redacted

Juan Pablo Vielma
Assistant Professor
Thesis Supervisor

Accepted by ................

Signature redacted

Dimitris Bertsimas
Boeing Professor of Operations Research
Co-Director, Operations Research Center
Predicting Performance Using Galvanic Skin Response

by

Lee Carter Mundell

Submitted to the Sloan School of Management
on May 11, 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

The rapid growth of the availability of wearable biosensors has created the opportunity for using physiological signals to measure worker performance. An important question is how to use such signals to not just measure, but actually predict worker performance on a task under stressful and potentially high risk conditions. Here we show that the biological signal known as galvanic skin response (GSR) allows such a prediction. We conduct an experiment where subjects answer arithmetic questions under low and high stress conditions while having their GSR monitored. Using only the GSR measured under low stress conditions, we are able to predict which subjects will perform well under high stress conditions with a median accuracy of 75%. If we try to make similar predictions without using any biometric signals, the median accuracy is 50%. Our results suggest that performance in high stress conditions can be predicted using signals obtained from GSR sensors in low stress conditions.

Thesis Supervisor: Tauhid Zaman
Title: Assistant Professor

Thesis Supervisor: Juan Pablo Vielma
Title: Assistant Professor
Acknowledgments

I wish to thank Professors Tauhid Zaman and Juan Pablo Vielma of the Massachusetts Institute of Technology’s Sloan School of Business. Their guidance and assistance throughout every stage of this thesis was invaluable. I also want to thank MIT’s Behavioral Research Lab for their assistance in recruiting subjects for and hosting the lab experiments used to develop this research. Finally, I would like to thank my finance, family, and friends for their support and patience with me for boring them with conversations about physiological signals, wearable devices, stress and productivity over many meals and coffee breaks in the past two years.
## Contents

1 **Introduction** 13
   1.1 Motivation for studying galvanic skin response 14
   1.2 GSR and stress detection 15
   1.3 Stress and performance 16
   1.4 Connecting GSR and performance 18

2 **Experiment** 19
   2.1 Experimental motivation 19
   2.2 Experimental design 20

3 **Exploratory Data Analysis** 25
   3.1 Statistics by round 25
   3.2 Performance features 26
   3.3 GSR features 28

4 **Statistical Modeling** 31
   4.1 Defining the classification task 31
   4.2 Modeling above median high stress earnings 33
   4.3 Modeling above 40th and 60th percentile high stress earnings 34
   4.4 Modeling continuous high stress earnings 35
   4.5 Modeling above median high stress accuracy 36

5 **Predictive Power** 39
   5.1 Methodology 39
6 Discussion
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Yerkes-Dodson Curve [27]</td>
</tr>
<tr>
<td>2-1</td>
<td>Hypothetical Yerkes-Dodson Curves.</td>
</tr>
<tr>
<td>2-2</td>
<td>a) Screenshot of Game Interface, b) Picture of NeuLog GSR Device, c) Round Order and Description</td>
</tr>
<tr>
<td>3-1</td>
<td>Mean Accuracy and GSR by Round.</td>
</tr>
<tr>
<td>3-2</td>
<td>Scatter Plot of Earnings, Calibration vs. High and Low vs. High.</td>
</tr>
<tr>
<td>3-3</td>
<td>Scatter Plot of Accuracy, Calibration vs. High and Low vs. High.</td>
</tr>
<tr>
<td>3-4</td>
<td>Examples of GSR Signal, Lowest and Highest Earners.</td>
</tr>
<tr>
<td>4-1</td>
<td>Scatter Plot of Above and Below Median Earners in 2-dimensional Feature Space</td>
</tr>
<tr>
<td>5-1</td>
<td>ROC Curves and AUC Values for Models from Table 4-1.</td>
</tr>
<tr>
<td>5-2</td>
<td>Boxplots of AUC Distribution for Above Median Earnings Models from Table 4-1.</td>
</tr>
<tr>
<td>5-3</td>
<td>Boxplots of AUC Distribution for Above 40th Percentile Earnings Models from Table 4-2.</td>
</tr>
<tr>
<td>5-4</td>
<td>Boxplots of AUC Distribution for Above 60th Percentile Earnings Models from Table 4-3.</td>
</tr>
<tr>
<td>5-5</td>
<td>Boxplots of AUC Distribution for Above Median Accuracy Models from Table 4-5.</td>
</tr>
</tbody>
</table>
6-1  a) GSR Signals Gathered From Two Devices Simultaneously b) The Ratio of the Two Signals.

47
List of Tables

3.1 Pearson Correlation Coefficients for Performance Metrics in Each Round. 28

4.1 Logistic Regression Models for Above Median Stress Earnings . . . . . 33
4.2 Logistic Regression Models for Above 40th Percentile high stress earnings 35
4.3 Logistic Regression Models for Above 60th Percentile high stress earnings 36
4.4 Linear Regression Models for High Stress Earnings, Continuous . . . 36
4.5 Logistic Regression Models for Above Median High Stress Accuracy . 37
Chapter 1

Introduction

For many jobs, it is difficult to test a worker’s performance under real-life conditions either because it is prohibitively expensive or because the stakes are too high to risk an error. For instance, it can be difficult to simulate the actual pressure and risk faced by a financial trader on a trading floor. The true risk of an accident is hard to recreate for air-traffic controllers, airline pilots, or truck drivers. A soldier can only experience the reality of combat on the battlefield. In each of these examples, the worker’s performance is evaluated using some sort of testing in a simulation or another low stress environment. For instance, airline pilots and soldiers are often tested in computer-generated simulations [16, 18]. The performance of a worker in a simulator or in another low stress environment may not be a good predictor of her performance in a high stress situation because it is not obvious how she will respond to the additional stress. Therefore, there could be immense value in being able to separate individuals with a high stress tolerance for a task from data gathered during low stress scenarios. In this study, we will describe the results of an experiment we conducted to test whether data on subjects’ galvanic skin response (GSR) could make this separation feasible.
1.1 Motivation for studying galvanic skin response

GSR, also known in the literature as electrodermal activity (EDA) or skin conductance (SC), is a measure of the electrical conductivity of a subject's skin. Changes in GSR are the result of activity in the eccrine sweat glands, the major sweat glands of the body, which are controlled by the autonomic nervous system (ANS) [21]. The ANS is the control system for much of the body's internal organs and bodily functions and it operates unconsciously without deliberate control. For this reason GSR is widely considered the most sensitive physiological indicator of psychological responses [30].

Interest in GSR has grown in recent years in part due to the growing market of wearable technology. In 2015, the estimated value of the global wearable devices market was $2 billion, up from less than $1 billion in 2012, and is predicted to triple to nearly $6 billion by 2018 [10]. Wearable device developers have begun to take advantage of the fact that the GSR signal is so easily obtained. A long list of new wristband devices have included sensors for GSR, including the Basis Peak [2], Neumitra [6], Microsoft Band 2 [5], and the Jawbone UP3 [4], among others.

Until recently, usage of GSR has been confined to lab experiments with limited practical application, with the exception of its key role in the well known polygraph or lie detector test [36]. A 1972 review of methods for detecting deception found GSR was far superior to all other variables tested [32]. More recently, GSR has been studied as a tool for emotional state recognition [29, 37], which has created opportunities for its use in behavioral research consulting [8], consumer insights [7], advertisement optimization [9], and affective computing [1]. However, few applications have been found for incorporating GSR monitoring into daily life as is now possible with the advent of more GSR-capable wearable devices. The only dedicated use for GSR that we have seen in a wearable is Empatica's Embrace, a wristband monitor that can predict seizures in epileptic subjects [3]. This relative void of practical applications for the new wealth of GSR data we can expect in coming years creates an opportunity for research connecting GSR to meaningful outcomes.
1.2 GSR and stress detection

Our study will focus on a potential use for GSR sensors that has drawn media interest with the release of new wearable devices: as a stress detection tool for the workplace [15, 22]. GSR has a long history of being studied as an indicator of stress. In 1907 Carl Jung conducted a series of therapy sessions with patients connected to an early GSR monitoring device[23]. In these sessions, patients were asked to participate in a word association game consisting of Jung prompting them with single word cues and asking them to respond quickly and honestly with another word they associated with the prompt. Jung noted that GSR levels spiked markedly for word associations attached to stressful or traumatic experiences in the subject’s lives [23].

More recently researchers have used lab experiments to develop statistical models to detect a change in stress levels experienced based on GSR data. In 2002, Andrew Lo and Dimitri Repin found that financial traders experience significant changes in GSR during stressful market volatility [25]. In 2005, Zhai et al. used GSR, heart rate, and pupil dilation to detect different stress states induced by a Stroop Test with up to 80% accuracy [40]. Also in 2005, Healey and Picard conducted a pioneer study of GSR in a real world setting. In their study, subjects’ physiological signals, including GSR, were monitored in three driving scenarios chosen to induce different levels of stress: at rest in a parking lot, highway driving, and city driving. They found that a recognition algorithm could correctly classify the different stress states with 97.4% accuracy, and that GSR features were particularly powerful for this task. Since then, many studies have found similar results detecting stress with above 90% accuracy both in real world settings [35] and in lab experiments with a variety of tasks and stress induction methods, including spatial and visual reasoning [34], hyperventilation and talk preparation [35], reading tasks [31], and arithmetic tasks [19]. In a 2010 study, Setz et al. were able to detect a stressed state with 83% accuracy using GSR features alone [33].

The above studies were motivated by an interest in developing systems to detect dangerous or unacceptable levels of stress in workers in a variety of different fields,
ranging from occupations involving driving and aviation to any occupation involving possibly stressful human computer interaction. All of the studies were successful in detecting stress, but none of them actually showed that the stress induced in their study was excessive or detrimental by connecting it to a metric of performance. For example, Healey and Picard show that city driving elicits a distinct and recognizable pattern in physiological signals compared to highway driving or at rest parking, but that does not by itself indicate whether a driver is under too much stress, making them prone to dangerous mistakes or behavior. In general, studies in this field of research implicitly assume that more stress is uniformly a bad thing. However, there is a significant field of research connecting stress and performance which reveals a more complicated relationship.

1.3 Stress and performance

The study of the relationship between stress and performance also began around the turn of the 20th century. In 1908, researchers Yerkes and Dodson conducted an experiment to study the relationship between learning rate and intensity of what they termed at the time "arousal" in mice [39]. The experiment consisted of running mice through repeated trials of a navigation decision where they had to choose between a dark and light colored room. When mice chose the dark colored room they received an electrical shock whereas choosing the light colored room incurred no punishment. In the study, Yerkes and Dodson varied the intensity of the electrical shock for different mice, and they found that the mice’s learning rates, the time it took for them to consistently choose the light colored room, followed a goldilocks pattern with respect to the shock intensity. At low levels of shock, the mice were not motivated enough to learn and often never learned the proper behavior. At the other extreme, excessively high voltage shocks were disruptive and led to decrements in the mouse’s learning rate. The fastest learning rates on average were achieved by medium intensity shocks. This finding has been canonized as the Yerkes-Dodson Law or Curve. Figure 1-1 shows a curve from a Forbes magazine article [27] illustrating how the relationship is often
characterized today.

The Yerkes-Dodson Law

![Yerkes-Dodson Curve](image)

Figure 1-1: Yerkes-Dodson Curve [27].

Unfortunately, revealing the Yerkes-Dodson Curve in human subjects has been more difficult. Many studies of stress or anxiety and performance in humans have only found a consistently negative relationship between stress and performance [38, 13]. A 1999 meta-study of the effect of financial incentives on performances in lab experiments found that in some cases incentives improved performance and in others decreased it [17]. The study attempted to distinguish these cases by task or stressor type, but in general no conclusive rules were found.

More recent studies of financial incentives and performance have found some evidence of a Yerkes-Dodson curve, but they have also noted that the effect of stress or incentives on performance is heterogeneous, i.e. some subjects are barely impacted by the stressors or positively motivated whereas others are seriously hindered [12, 14]. In the framework of the Yerkes-Dodson curve, this heterogeneity could be explained by differences in the shape of individual’s curves, a possibility we explore in Chapter 2 when explaining the motivation for our experiment.
1.4 Connecting GSR and performance

We hope to use our study to contribute to bridging the gap between the two fields of research, GSR and stress detection and stress and performance, by connecting GSR directly to performance outcomes. Only a few papers have attempted to explore the space of physiological signals and performance. Two studies considering the effects of incentives and stressors on performance found recognizable patterns in the neural activity captured with fMRIs of high performers versus low performers [28, 26]. Another active area of research has connected testosterone and cortisol levels with financial risk-taking and performance [20, 11]. However, in both of these cases the data points are difficult and invasive to acquire, a significant obstacle for real-time practical applications. Only one study, conducted by Lin et al. in 2005, connected GSR with performance [24]. In their experiment, subjects played a Mario video game while connected to GSR sensors. They found that strongest performers in the game had low normalized GSR levels on average and the poorest performers had high mean normalized GSR levels on average. All of these studies found concurrent relationships between physiological signals and performance, not a predictive one. We hope to go a step further in our study by showing that more than just correlating reliably with performance, GSR features can be used to predict performance under higher stress conditions. A truly predictive relationship would significantly increase the range of practical applications for GSR monitoring in the real world.

The rest of this thesis will be organized as follows. Chapter 2 explores the motivation for and design of our human subjects experiment. Chapter 3 describes the data collected from the experiment and the features that we developed from that data. Chapter 4 describes the statistical models we developed for predicting performance outcomes. Chapter 5 illustrates the predictive power of our models. Finally, Chapter 6 concludes the thesis with discussion and thoughts on future work.
Chapter 2

Experiment

2.1 Experimental motivation

We designed a lab experiment with human subjects with the goal of uncovering a predictable relationship between GSR and actual performance. Although an ideal experiment might attempt to capture and map out the entire Yerkes-Dodson Curve for each participant, we decided for both practical and theoretical reasons that it was not necessary. In order to map out a Yerkes-Dodson Curve for a subject, our experiment would need to expose them to many carefully calibrated treatments designed to induce incrementally varied levels of stress. Reviewing the literature on inducing stress in a lab setting, it is clear that none of the existing methods provide sufficient granularity of stress manipulation—human subjects cannot be motivated by electric shocks as Yerkes and Dodson did to mice in their original study. Moreover, knowing an individual’s full Yerkes-Dodson Curve is not especially useful in application. Real world situations rarely offer the ability to incrementally reduce or increase stress as would be necessary to optimize a task for a given Yerkes-Dodson Curve. Instead, the types of decisions we believe that GSR can help with are more often binary: is this trainee sufficiently comfortable with a task to perform well under pressure? Or from the perspective of a hiring manager: which candidates will perform the required task well under pressure? Figure 2-1 illustrates this scenario with a series of hypothetical Yerkes-Dodson Curves for different individuals. In the figure, individuals 1 and 2
Figure 2-1: Hypothetical Yerkes-Dodson Curves.

have curves such that, at a hypothetical stress level associated with a task marked by the solid line, they would perform significantly worse than individuals 3 and 4. The goal of our experiment is not to trace out the entire curve for each individual, but rather to identify whether we can separate in advance the individuals with higher performance under stress, 3 and 4, from individuals who choke under pressure, 1 and 2.

2.2 Experimental design

For our study, we recruited 30 participants who were required to be undergraduate students enrolled in a Boston area university. Our participants were recruited through the MIT Behavioral Research Lab where the trials were conducted. The subjects were told they would be participating in a study of mathematical ability and stress and signed a consent form. Following the trial, all subject data was de-identified.

Each subject was fitted with a NeuLog GSR finger sensor on the index and middle fingers on their non-dominant hand. A picture of the device and sensor apparatus is shown in Figure 2-2a. The GSR signal was captured throughout the whole experiment at a 5 Hz sampling rate. Early trials of the study were run using a wearable wristband
device, the Empatica E4. However these trials had to be thrown out because of the number of times the signal was lost, difficulties with the software, malfunctioning hardware parts, etc. The device is marketed as a "clinical grade" $1,500 dollar research device, so we would like to make the caveat that even the state of the art wearable devices may not yet have completely solved the difficulties of reliably capturing GSR. The experiment consisted of multiple stages with varying degrees of stress but uniform task difficulty. The subject was first asked to relax for three minutes while we ensured that the monitoring device was working. After this rest period, they played three distinct rounds of a computer-based arithmetic game consisting of a series of three digit by two digit multiplication questions. A screenshot of the game interface is shown in Figure 2-2b. Between each round the subject was given a three minute rest period. The ordering of rounds and rest periods is visualized in Figure 2-2c.

The first round of the game was used for calibration and designed to be minimally
stressful in order to evaluate the subjects’ baseline performance and GSR response to the arithmetic questions. Subjects answered as many multiplication questions as possible in ten minutes. They did not receive any feedback on whether or not their answers were correct and there was no time limit for answering individual questions (the timer bar shown in Figure 2-2b was only visible in later rounds). During this round, subjects were motivated by a small financial reward of $0.25 for each correct response. This round provided us with subjects’ physiological stress to the arithmetic task without significant stress. During the rest period following the calibration round, the subject’s average response time was calculated and used in the following two rounds for the incentive scheme described below.

The second and third rounds of the game were the low and high stress rounds. Both of these rounds consisted of 20 multiplication questions. Subjects were limited to 130% of their average answer time from the calibration round to answer each question in these rounds. A portion of the time allowed was "bonus" time, during which correct answers were worth $1 instead of $0.25. During the low stress round, 85% of the total question time allotted was bonus time, but during the high stress round only 50% of the time was bonus time. These percentages were chosen such that achieving the bonus payout in the low stress round was only slightly challenging, whereas in the high stress round it was very difficult. The time remaining for each question was visible to the subjects on the game screen as a bar that reduced in size every second, with bonus and normal time represented by green and red shading respectively, as shown in Figure 2-2b. Subjects were alerted when they ran out of bonus time by an unpleasant buzzer sound. During these two stress rounds, subjects were also given feedback indicating whether the answer they submitted for each question was correct.

To summarize, the first round was designed as a baseline for capturing GSR and performance on the tasks with little to no stress involved. For the final two rounds, three forms of stress were added: financial stress induced by the tiered payout structure and the potential to make significantly more money, timing stress induced by the timer bar and bonus time, and a degree of social pressure induced by the feedback which was visible to both the subject and the researcher sitting next to them.
The significantly more challenging timing scheme in the third, high-stress round was designed to induce more stress than the second, low-stress round.
Chapter 3

Exploratory Data Analysis

3.1 Statistics by round

Two different outcome metrics can be derived from our experiment: accuracy on the questions asked and monetary earnings in each round. In our analysis, we focus on the earnings metric as the outcome of interest because earnings is likely to be more sensitive than accuracy to the impact of stress on the subject’s performance. We include analysis showing that the accuracy is not well predicted by either performance features or GSR features. Earnings depends upon both how fast the subject answers each questions and also how many questions are answered correctly, whereas the accuracy just depends upon the number of questions answered correctly. If a subject is impacted by the stress negatively, this will have a larger impact on their earnings (which is sensitive to speed) than accuracy (which can remain high even if the subject slows down). If we believe that GSR is related to the stress of the subject, and that the stress is impacting performance, then we would expect the good and bad performing subjects to be more easily distinguished in earnings than in accuracy.

Figure 3-1 shows the mean accuracy and mean GSR across all participants in each round—for each subject, the time-averaged GSR level is calculated within each round and the mean of these averages is shown. Earnings cannot be compared across rounds because of the change in payout structure, but this does not pose a problem for our main goal since we are interested in distinguishing performers within rounds
rather than between rounds. Unfortunately, due to limited sample size none of the
differences in the means shown are statistically significant, so the trends are merely
suggestive. However, we see in the accuracy plot that performance seems to follow a
Yerkes-Dodson Curve; and in the GSR plot, we see that GSR, and therefore stress,
increases on average between rounds.

3.2 Performance features

The default hypothesis for our study is that performance is correlated across rounds,
regardless of stress: subjects who perform well in the calibration and low stress rounds
will also perform well in the high stress round. If this is the case, then performance
features from the calibration and low stress rounds should be highly correlated with
earnings in the high stress round, and therefore GSR would be unnecessary and ir-
relevant for identifying high performers. However, we find that these features are
not highly correlated with either metric of performance, earnings or accuracy, in the
higher stress rounds. Figures 3-2 and 3-3 plot earnings and accuracy respectively for
each subject in calibration and low stress rounds versus the high stress round. We
can see that no obvious relationship exists between the performance metrics across
rounds. The lack of a significant relationship between performance across periods is
confirmed by the correlation coefficients for these metrics, which are not statistically significant at the five percent level for each performance metric. Table 3-1 shows the correlations between all of the performance metrics, including average response time in each round. The portions of the table shaded in gray indicate the pairs we are most interested in. No significant correlations exist between either calibration round performance and high stress performance or low stress performance and high stress performance, suggesting that those features would likely be poor predictors of high
stress performance outcomes. This supports our hypothesis that to predict performance under high stress conditions, we will likely need more information than simply performance in lower stress conditions, and also our speculation that simulation and testing methods designed to measure a candidate's ability in lower stress conditions are not necessarily good indicators of real-world performance by themselves. Table 3-1 also demonstrates that average answer times in the calibration and low stress rounds are not correlated with outcomes, indicating that our time limit calibration scheme is not unfairly benefiting fast or slow subjects.

Table 3.1: Pearson Correlation Coefficients for Performance Metrics in Each Round.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Calibration Accuracy</th>
<th>Calibration Earnings</th>
<th>Calibration Answer Time</th>
<th>Low Stress Accuracy</th>
<th>Low Stress Earnings</th>
<th>Low Stress Answer Time</th>
<th>High Stress Accuracy</th>
<th>High Stress Earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration Accuracy</td>
<td>0.63***</td>
<td>-0.45*</td>
<td>0.40*</td>
<td>0.367***</td>
<td>0.50**</td>
<td>0.32</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Calibration Earnings</td>
<td>-0.45**</td>
<td>-0.87***</td>
<td>0.18</td>
<td>0.18</td>
<td>-0.26</td>
<td>-0.12</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Calibration Answer Time</td>
<td>-0.45*</td>
<td>-0.87***</td>
<td>0.18</td>
<td>0.18</td>
<td>-0.26</td>
<td>-0.12</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Low Stress Accuracy</td>
<td>0.40*</td>
<td>0.18</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.02</td>
<td>0.97***</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>Low Stress Earnings</td>
<td>0.56</td>
<td>0.05</td>
<td>-0.02</td>
<td>0.96</td>
<td>-0.02</td>
<td>0.28</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Low Stress Answer Time</td>
<td>-0.45*</td>
<td>-0.87***</td>
<td>-0.14</td>
<td>-0.14</td>
<td>-0.02</td>
<td>-0.27</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>High Stress Accuracy</td>
<td>0.32</td>
<td>0.18</td>
<td>-0.34</td>
<td>0.33</td>
<td>0.28</td>
<td>-0.02</td>
<td>0.61***</td>
<td></td>
</tr>
<tr>
<td>High Stress Earnings</td>
<td>0.26</td>
<td>-0.12</td>
<td>0.06</td>
<td>0.18</td>
<td>0.20</td>
<td>-0.01</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

3.3 GSR features

Next, we consider the predictive power of the GSR signal we measured for each subject. Figure 3-4 shows the trajectory of the raw GSR signal over the course of the experiment for two subjects, who in this case are the highest and lowest earners in the high stress round. In the figure, each round of the trial is separated by a short rest period, demarcated with red lines, and each round of the experiment is labeled. The GSR signal is measured in microsiemens (µS), a unit of conductance. The figure shows that the GSR signal can exhibit a variety of behaviors, such as upward and downward trends and rapid oscillations. There are also significant differences in the GSR signal path for these two subjects. A common criticism of GSR in the literature we reviewed is that the signal is too heterogeneous across individuals and that this reduces the usefulness and increases the difficulty of working with GSR. However, we believe this heterogeneity may itself contain useful value for distinguishing performers,
Figure 3-4: Examples of GSR Signal, Lowest and Highest Earners.

and we will show later that in fact our most predictive features are not normalized by subject.

Our goal is to create features from the GSR signal that will predict performance in the high stress round. The GSR signal is very high dimensional, with over 10,000 sampled values per subject. To obtain a useful characterization the GSR signal for each user, we reduced the full high dimensional signal to a low dimensional set of features. In our exploratory analysis, several of the features we tested did not prove to be significant predictors. We include a description here of all features we considered:

**Drift**: The change in GSR from the beginning to the end of a round

**Normalized Drift**: The change in GSR from the beginning to the end of a round divided by the time averaged GSR signal during that round

**Average**: The time averaged level of GSR during a round

**Shift**: The percentage change in the time averaged GSR signal in the current round and the previous round. (Shift requires a previous round for reference, so it is
only defined for the low stress and high stress rounds, not the calibration round)

**Maximum Increase:** The increase in GSR from its minimum point during a round to the end of the round.

**Maximum Increase Slope:** The increase in GSR from its minimum point during a round to the end of the round divided by the time interval between those two points.

**Normalized Maximum Increase Slope:** The increase in GSR from its minimum point during a round to the end of the round divided by the time interval between those two points and the time averaged level of GSR during the round.

Our final GSR feature models only used two features engineered from the GSR signal, drift and maximum increase. Intuitively, drift measures slow variations in the GSR signal over the course of a round and maximum increase measures identifies trends in the GSR signal that are present at the end of the round. A 10,000-dimensional space offers far more complex and creative features than we explored here. However we believe that part of the significance of our findings is that simple time series features like the ones described prove to be valuable for prediction.
Chapter 4

Statistical Modeling

4.1 Defining the classification task

Our classification task is to distinguish between subjects who perform well in the high stress round from those who do not using data from the calibration and low stress rounds. For this purpose, we assign a binary label to each subject indicating whether he/she is a good or bad performer in the high stress round. Our simple labeling scheme classifies a subject as a good performer if their performance metric is strictly above the median value for all subjects, otherwise the subject is a bad performer. We use the median performance metric as a threshold because it is robust to outliers with extreme values for the metric.

To gain a better understanding of our data, we show a scatter plot of the two classes, good and bad performers, in a two dimensional feature space in Figure 4-1. In the scatter plot the good and bad performers are indicated with different shaped markers. We chose the features drift and maximum increase from the calibration to demonstrate just how well the GSR data separates classes with only two features.

It is not immediately clear why these features are effective at separating the good and bad performers. From the figure, the bad performers typically have a low maximum increase and a drift that is near zero or negative in the calibration round. This suggests that, for bad performers, the GSR signal does not increase very much during the calibration round. If we accept the standard theories in the GSR literature that
suggest that increases in GSR indicate an increase in stress or cognitive load, then the bad performers are not suffering very much stress or cognitive load in the calibration round. In contrast, the good performers have a large maximum increase relative to their drift. These subjects are experiencing a higher level of stress or cognitive load. One possible explanation is that bad performers do not become focused in the absence of any stress or external motivator, and then when the stress is added they are not prepared to perform well. The good performers become mentally focused naturally without any additional external stressors. Then, when external stress is added, they are not affected by it because they have already prepared themselves mentally. This is only one possible explanation for what we observe and there may be other explanations, but we do not explore them further. As operations researchers, our goal is not to provide an intuitive explanation of the relationship between GSR and performance, but rather to predict who will perform well under stress.
4.2 Modeling above median high stress earnings

To evaluate the statistical significance of these apparent relationships, we developed logistic regression models for the above median high stress earnings class. For the performance features, we considered accuracy, earnings and average answer time. For the GSR features, we considered all of the features described in Chapter 3, from both calibration and low stress rounds. Table 4-1 shows estimation results for models from

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Performance Model #1</th>
<th>Performance Model #2</th>
<th>GSR Model #1</th>
<th>GSR Model #2</th>
<th>Combined Model #1</th>
<th>Combined Model #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration Accuracy</td>
<td>7.70*</td>
<td>11.08*</td>
<td>6.66</td>
<td>8.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.86)</td>
<td>(4.94)</td>
<td>(3.95)</td>
<td>(4.98)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Answer Time</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Earnings</td>
<td>-0.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.38)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration GSR Drift</td>
<td>-1.39*</td>
<td></td>
<td>-1.47*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.65)</td>
<td></td>
<td>(0.71)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration GSR Max. Increase</td>
<td>2.82**</td>
<td></td>
<td>3.01**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.02)</td>
<td></td>
<td>(1.09)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Stress GSR Drift</td>
<td></td>
<td>-1.20</td>
<td>-1.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.70)</td>
<td>(0.71)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Stress GSR Max. Increase</td>
<td>3.27*</td>
<td></td>
<td>4.01*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.33)</td>
<td></td>
<td>(1.65)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Akaike Information Criterion</td>
<td>42.17</td>
<td>40.04</td>
<td>35.56</td>
<td>36.92</td>
<td>33.62</td>
<td>34.33</td>
</tr>
</tbody>
</table>

(Std. Error), *: p < 0.05, **: p < 0.01

Table 4.1: Logistic Regression Models for Above Median Stress Earnings

three different feature sets: performance features, GSR features, and performance and GSR features combined. We compared models according to feature significance and Akaike Information Criterion (AIC) and report in the table the two strongest models from each feature set. AIC is a measure of quality of a statistical model that considers goodness of fit while penalizing excess complexity. A lower AIC represents a superior quality model. We see that the models built with performance features alone contain features from the calibration round only. No performance features from the low stress round were found to be significant predictors of high stress earnings. The best performance feature models contain only one weakly significant predictors each. The two best GSR feature models both pair drift and maximum increase features from the same period. The magnitudes of the coefficients for drift and maximum increase
in the GSR feature models are roughly the same, suggesting a consistent relationship between these features in lower stress states and the high stress state. The calibration round model exhibits both greater statistical significance for both predictors and a lower AIC than the low stress round model. Finally, the best combined feature models use the same features as the GSR feature models with the addition of the calibration accuracy feature, which is not statistically significant in either model. However, the AIC for the combined models does improve slightly over the GSR feature models. With that caveat, it appears that the GSR feature models have the best fit for the data and little is achieved by additional complexity or by adding performance features.

4.3 Modeling above 40th and 60th percentile high stress earnings

We performed the same analysis for two additional classification thresholds separating good and bad performers: 40th percentile and 60th percentile. The results are shown in Tables 4-2 and 4-3. We can see in Table 4-2, showing the 40th percentile classifiers, that the GSR feature models achieve significantly lower AIC than the performance models, and the combined models again achieve only slight improvements in AIC over the GSR models, corroborating our findings with the median threshold classifiers. Table 4-3, showing the 60th percentile classifiers, tells a slightly different story. In this case, the second performance model is very powerful and the best combined model includes two performance features and only one GSR feature, which is not significant. One possible explanation for this is that the calibration earnings and accuracy features used in the performance model are significantly collinear (see Table 3-1). The fact that their coefficients are of opposite signs and so much larger than they were in our previous models suggests that we should be suspicious of those models and their significance. Alternatively, the relative weakness of the GSR features compared to the 40th percentile classifiers suggests that perhaps GSR is not as powerful for separating out the highest performers (top 40%) from the pack, but rather more
useful for separating out especially low performers (bottom 40%) instead.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Performance Model #1</th>
<th>Performance Model #2</th>
<th>GSR Model #1</th>
<th>GSR Model #2</th>
<th>Combined Model #1</th>
<th>Combined Model #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration Accuracy</td>
<td>7.94* (3.74)</td>
<td>10.69* (4.63)</td>
<td></td>
<td></td>
<td>9.77 (5.92)</td>
<td>6.50 (3.86)</td>
</tr>
<tr>
<td>Calibration Answer Time</td>
<td>0.05 (0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Earnings</td>
<td>-0.65 (0.38)</td>
<td>-0.89 (0.64)</td>
<td>-1.19 (0.74)</td>
<td>5.36* (2.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration GSR Drift</td>
<td></td>
<td>4.60** (1.79)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration GSR Max. Increase</td>
<td>-1.51 (0.81)</td>
<td>-1.52 (0.88)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Stress GSR Drift</td>
<td>3.38* (1.46)</td>
<td>3.93* (1.69)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Stress GSR Max. Increase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Akaike Information Criterion</td>
<td>40.28</td>
<td>38.76</td>
<td>28.38</td>
<td>35.66</td>
<td>25.63</td>
<td>33.67</td>
</tr>
</tbody>
</table>

Table 4.2: Logistic Regression Models for Above 40th Percentile high stress earnings

### 4.4 Modeling continuous high stress earnings

Although our main analysis focuses on GSR's usefulness for separating low and high performer classes, we also tested GSR's predictive power for the continuous outcome of high stress earnings using linear regression. We used the same methodology described for the binary classification feature selection, selecting the models with the best performance using features from the same three sets. In this case, we evaluated models according to coefficient significance and improvements in adjusted r-squared. Our best models for this task are shown in Table 4-4. None of the features from the low stress round, either performance or GSR, made it into the best models in this analysis. The first, single-variable GSR model provides a dramatic improvement in adjusted r-squared over the best performance feature model. Based on the adjusted r-square values, performance features alone explain essentially none of the variation in high stress earnings. The final combined model shows that when GSR is included nearly a third of the variation can be explained and the GSR features are highly
Table 4.3: Logistic Regression Models for Above 60th Percentile high stress earnings significant.

Table 4.4: Linear Regression Models for High Stress Earnings, Continuous

4.5 Modeling above median high stress accuracy

Finally, we perform the classification analysis for accuracy in the high stress round rather than earnings to test our assumption that accuracy would show very little
predictable variation and is a poor outcome metric for our study. Table 4-5 shows the best logistic regression models we were able to develop for classifying above median high stress accuracy over the same three feature sets. The table illustrates that we were unable to find any models with significant predictors when either only performance features or GSR features were used. When we combine the feature sets, our final model has one significant predictor but we believe that this is a result of collinearity and overfitting as AIC does not improve significantly from any of the other models.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Performance Model #1</th>
<th>GSR Model #1</th>
<th>Combined Model #1</th>
<th>Combined Model #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration Accuracy</td>
<td>3.57</td>
<td>4.24</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td>Calibration Accuracy</td>
<td>(3.22)</td>
<td>(2.90)</td>
<td>(3.17)</td>
<td></td>
</tr>
<tr>
<td>Calibration Answer Time</td>
<td></td>
<td></td>
<td>-0.06</td>
<td></td>
</tr>
<tr>
<td>Calibration GSR Drift</td>
<td></td>
<td></td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Calibration GSR Max. Increase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Stress Accuracy</td>
<td>3.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.61)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Akaike Information Criterion</td>
<td>43.69</td>
<td>43.90</td>
<td>43.49</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.5: Logistic Regression Models for Above Median High Stress Accuracy
Chapter 5

Predictive Power

5.1 Methodology

To estimate how much of an improvement GSR features offer over performance features for classifying good and bad performers, we evaluate the prediction accuracy our models from Chapter 4 using stratified, four-fold cross-validation. This technique involves separating the data into four separate sets or folds with balanced numbers of each outcome class in all folds. The model being tested is fit on three of the four folds and predictions are generated on the fourth held-out fold. This fit and predict process is repeated four times, with a different fold being held out in order to generate a prediction for each data point. These predictions are then the basis for evaluating the power and accuracy of the model. We use the holdout predictions to develop receiver operating characteristic (ROC) curve, a visualization of the trade-off between the true positive and false positive rates of your predictions given different values for the decision rule. Since the predictions from a logistic regression are continuous between 0 and 1, we can consider different thresholds for classifying a prediction as a 1 or a 0. The ROC curve is drawn by finding each threshold that leads to a change in the class predictions and then plotting the true positive rate and false positive rate at each of those thresholds. With the complete ROC curve, we can evaluate a model’s predictive power by calculating the area under the curve (AUC). An AUC of 0.5 indicates that a predictive model is no better than random guessing and the closer an AUC is to
1.0, the more accurate it is considered to be. Using AUC to evaluate a model is more powerful than only considering accuracy, the percentage of all classifications that are correct, at a single threshold, typically 0.5. This is especially true in our case because our goal is to separate and rank good and bad performers, not to achieve a specific prediction value, greater than or less than 0.5.

Figure 5-1 shows three ROC curves and notes the corresponding AUCs for three of the models from Table 4-1.

![ROC Curves and AUC Values for Models from Table 4-1.](image)

One problem with this analysis is that our results may be sensitive to the splits chosen for the 4-fold cross-validation. To test the robustness of our models to this element of randomness, we generated 500 different permutations of stratified folds to evaluate each model. Figure 5-2 shows boxplots for the distribution of AUCs for each of the six models from Table 4-1. Note from the figure, the performance feature models perform no better than random guessing (AUC = 0.5) whereas the GSR feature models achieve a median AUC of 0.76. The combined feature models do not appear to achieve higher AUC than the GSR feature models, suggesting that performance features provide very little if any predictive information not already contained in the GSR features for our data. We can also evaluate these models for the accuracy metric described above instead of AUC. The result is a median accuracy
Figure 5-2: Boxplots of AUC Distribution for Above Median Earnings Models from Table 4-1.

of 75% for both the best GSR model and the best combined model and a median accuracy of only 50% for the performance feature models.

We also considered the predictive power of modeling the two alternative classification thresholds separating good and bad performers: 40th percentile and 60th percentile. Figures 5-3 and 5-4 show the corresponding boxplots for those analyses using the models from Tables 4-2 and 4-3. Figure 4-3 shows the same story that

Figure 5-3: Boxplots of AUC Distribution for Above 40th Percentile Earnings Models from Table 4-2.
we saw in our main analysis. The performance feature models often underperform random guessing whereas the GSR models offer a significant improvement. Although combining the two feature sets improved our models in terms of significance and AIC, the actual predictive improvement is approximately zero. Figure 4-4 verifies our concern from the Table 4-3 that the performance model which appears to perform so well, column 2, actually exhibits huge variance and has a median significantly below the GSR models. Moreover, the combined model that attempts to leverage both features from that performance model, column 6, also underperforms the GSR models. We can safely say that the significance found in those performance models was spurious and that in fact or basic story remains the same: GSR features outperform performance features in terms of predictive power.

We also ran the same procedure for above- and below-median accuracy labels to verify that that outcome metric is in fact less predictable and hard to distinguish than the earnings outcome. Figure 5-5 shows the predictive accuracy of the 4 models from Table 4-5. The figure shows that these accuracy models are more variable with respect to the training data permutation (wider spread in the boxplot) and do not achieve nearly as high median AUC when compared with the earnings models. Moreover, the GSR feature model does not significantly outperform the performance feature model.
Figure 5-5: Boxplots of AUC Distribution for Above Median Accuracy Models from Table 4-5.
Chapter 6

Discussion

Our findings suggest that GSR features are significant predictors of performance. Previous research in related fields demonstrated separately that GSR is a strong indicator of stress and that stress has an impact on performance, but our results link these findings by showing that GSR can be related to performance directly. More specifically, our findings demonstrate that GSR features captured in an unstressed state can be used to separate subjects who will perform poorly under high stress conditions from subjects who will perform well. This finding widens the scope of applications in which the GSR signal could be useful. If we had found that GSR was contemporaneously correlated with performance, that relationship would be of limited use given that by the time the model identified the critical GSR pattern the decline in performance would have already occurred. Our findings instead suggest that it is possible to predict in advance who is likely to under-perform and thereby take steps to prevent costly negative outcomes.

We believe these findings may have the largest direct impact in industries where significant training and simulation is used to prepare and evaluate candidates prior to exposure to the harsher conditions of the actual job, such as military and combat training, pilot training, law enforcement training, and emergency response training. In these fields, simulation and test results are currently the primary indicators of trainee preparedness. Our findings demonstrated that GSR features may be stronger predictors in these cases than those traditional performance outcomes. GSR could
be used as an additional tool, improving on the current methods, for gauging the preparedness of trainees for the rigors of their future responsibilities and help flag individuals who are not yet ready for those roles.

We also believe our study may serve as the motivation for future work looking deeper into the relationship between GSR and performance. Much of the interest in wearable devices and GSR derives from the potential for GSR to alert subjects to excessive stress which could lead to mistakes and a decline in performance. Our study helps to separate subjects whose performance is likely to decline soonest, but does not directly address the problem of predicting a decline in a subject before it happens. A study monitoring GSR over a longer time period could attempt to recognize patterns in physiological signals preceding decrements in performance throughout the day or week of an employee.

There are a number of caveats that should be considered when interpreting our findings. First, our sample size of 30 subjects limited both the tools of analysis at our disposal and the strength of our findings. A much larger study could leverage machine learning tools to explore the feature space of high-dimensional GSR data much more effectively and potentially find even more accurate methods of prediction.

Second, our experiment tests one particular type of task, arithmetic. The literature on stress and performance includes a wide range of tasks involving complete different types of brain function including visual spatial reasoning, reading and writing, etc..

Third, because of difficulties with the wearable wristband sensor devices that we purchased during the design phase of our experiment, discussed in Chapter 2, we were not able to run the experiment with a wearable wristband. Ideally, we would have liked to show that our result could have been achieved with the quality of signal collected by a non-invasive, inexpensive wearable. A recently released wearable device, the Microsoft Band 2, included significant improvements in hardware design which solved several of the issues we encountered with other wearables. Preliminary results, shown in Figure 6-1, comparing the signals from the Microsoft Band 2 and the NeuLog finger sensors suggest that the wearable technology may now be sufficiently accurate.
to replace less convenient finger sensors, at least for the models used in our study.

Figure 6-1: a) GSR Signals Gathered From Two Devices Simultaneously b) The Ratio of the Two Signals.

The first plot in the figure shows the two raw signals gathered during a 10-minute interval. The difference in the absolute level of the two signals is a result of the different locations and size of the sensors, fingertips for the NeuLog device and wrist for the Microsoft Band 2. We can see that the Microsoft Band 2's sensors are not sensitive enough capture some of the short-term fluctuations captured by the NeuLog sensors. Short-term fluctuations may not be important though. The second plot in the figure shows the ratio of the two signals, which remains relatively constant overall. A constant ratio between the two device signals suggests that they may be equally effective at capturing longer term trends in the signal like the time series features calculated over from 10-30 minute intervals in our study. Future studies should consider using newer wearable devices, like the Microsoft Band 2, since they may be sufficiently accurate to measure the kinds of features used in our study.

Finally, as our findings are the result of a lab experiment, they are not guaranteed to have external validity. Multiplying three by two digit numbers in a computer-based game likely requires some of the same mental resources necessary for more complex tasks, but it is yet to be proved that these relationships will hold in real world settings.

We hope to address each of these issues in our ongoing research into the practical applications for physiological data gathered from wearable devices.
Bibliography


[27] Stephen Meyer. Why the u.s. keeps losing the ryder cup: Research on stress provides answer.


