STRESS DETECTION FOR KEYSTROKE DYNAMICS

SHING-HON LAU

May 2018
CMU-ML-18-104

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee

Roy Maxion, Chair
Tom Mitchell
Dan Siewiorek
Peter Strick (University of Pittsburgh)
David Banks (Duke University)
Mark Wetherell (Northumbria University)

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

Abstract

Background. Stress can profoundly affect human behavior. Critical-infrastructure operators (e.g., at nuclear power plants) may make more errors when overstressed; malicious insiders may experience stress while engaging in rogue behavior; and chronic stress has deleterious effects on mental and physical health. If stress could be detected unobtrusively, without requiring special equipment, remedies to these situations could be undertaken. In this study a common computer keyboard and everyday typing are the primary instruments for detecting stress.

Aim. The goal of this dissertation is to detect stress via keystroke dynamics – the analysis of a user's typing rhythms – and to detect the changes to those rhythms concomitant with stress. Additionally, we pinpoint markers for stress (e.g., a 10% increase in typing speed), analogous to the antigens used as markers for blood type. We seek markers that are universal across all typists, as well as markers that apply only to groups or clusters of typists, or even only to individual typists.

Data. Five types of data were collected from 116 subjects: (1) demographic data, which can reveal factors (e.g., gender) that influence subjects' reactions to stress; (2) psychological data, which capture a subject's general susceptibility to stress and anxiety, as well as his/her current stress state; (3) physiological data (e.g., heart-rate variability and blood pressure) that permit an objective and independent assessment of a subject's stress level; (4) self-report data, consisting of subjective self-reports regarding the subject's stress, anxiety, and workload levels; and (5) typing data from subjects, in both neutral and stressed states, measured in terms of keystroke timings – hold and latency times – and typographical errors. Differences in typing rhythms between neutral and stressed states were examined to seek specific markers for stress.

Method. An ABA, single-subject design was used, in which subjects act as their own controls. Each subject provided 80 typing samples in each of three conditions: (A) baseline/neutral, (B) induced stress, and (A) post-stress return/recovery-to-baseline. Physiological measures were analyzed to ascertain the subject's stress level when providing each sample. Typing data were analyzed, using a variety of statistical and machine learning techniques, to elucidate markers of stress. Clustering techniques (e.g., *K*-means) were also employed to detect groups of users whose responses to stress are similar.

Results. Our stressor paradigm was effective for all 116 subjects, as confirmed through analysis of physiological and self-report data. We were able to identify markers for stress within each subject; i.e., we can discriminate between neutral and stressed typing when examining any subject individually. However, despite our best attempts, and the use of state-of-the-art machine learning techniques, we were not able to identify universal markers for stress, across subjects, nor were we able to identify clusters of subjects whose stress responses were similar. Subjects' stress responses, in typing data, appear to be highly individualized. Consequently, effective deployment in a real-world environment may require an approach similar to that taken in personalized medicine.

In memory of Martin Azizyan

and

To Ada, with lots of hugs and love

Acknowledgments

To the thesis committee (Roy, Mark, David, Tom, Peter, and Dan), thank you for your invaluable advice and assistance throughout my thesis project. I would not have been able to complete this project without your help.

To Pat, thank you for your tireless dedication to running subjects and collecting data for this thesis. Without you, I would not have had any data to analyze.

To Huayun, thank you for your help with writing software for this project and for your help in developing some of the visualizations in this thesis.

To my friends, Qirong, Nicole, Carlton, and especially Ada, thank you for your support throughout my PhD. I would not have made it through without you.

Contents

Chapter 1

Introduction

Stress is a common, familiar, and pervasive phenomenon. While typically viewed as a nuisance at worst and a good motivator at best, stress can be used as an indicator for a wide variety of phenomena. This is best illustrated by considering three different people: an air traffic controller, a disgruntled employee at a top-secret government facility, and an average office worker.

An air traffic controller is an example of an operator of critical infrastructure; other examples include electricity grid supervisors, nuclear power plant operators, and triage nurses in a hospital emergency room. Such operators are responsible for making vital decisions on a day-to-day, and often minute-to-minute, basis. Errors in decision making can lead to significant monetary loss, serious injury, or even death. Such catastrophic errors are more likely to occur when operators are under extreme stress. If it were possible to detect stress by analyzing the way operators interact with their computer systems, it would be possible to take counter-measures to reduce the likelihood of such errors occurring or to at least mitigate the impact of these errors. For example, an overstressed operator could be given additional support from on-call operators or be temporarily removed from duty until they regain their composure. Considering that errors may result in billions of dollars of damages and the loss of hundreds, if not thousands, of lives, any technique for reducing their prevalence is worth pursuing.

Employees at top-secret government facilities are responsible for tasks that are vital to national security. By virtue of their position, such employees necessarily have access to sensitive data and sensitive computer systems that are vital to the national interest. If such an employee becomes disgruntled and seeks to abuse their position to commit misdeeds, significant damage can be done to the country's interest. Malicious employees are often referred to as "insiders"; a prominent example of an insider in recent years would be that of Edward Snowden. It is easy to imagine that insiders are likely to be under considerable stress as they perpetrate their misdeeds. If this stress could be detected, immediate action could be taken. Security officers could be notified to detain the insider or temporary restrictions on data access could be imposed. Even the existence of stress-detection technology may deter future disgruntled employees from becoming insiders.

Finally, consider the average office worker. Such an individual is charged with far less responsibility than an operator of critical infrastructure or an employee working at a top-secret government facility, but such individuals form a large percentage of the world's population. Chronic stress, whether caused by the job itself, finances, home-life, or otherwise, can cause physical and mental health problems. This results in a lower quality-of-life and also contributes to decreased productivity at work. Such stress may not be obvious to the individual or even to those around him/her,

as its accumulation is often a slow, gradual process. However, if detection of stress were possible, individuals could have their levels of stress tracked over long periods of time without requiring any change in daily patterns. This could be done either at a personal level (e.g., smartphone app) or on an organizational level (e.g., monitoring software on every corporate workstation). Individuals could be made aware of their chronically elevated levels of stress, permitting remedial actions to be taken. Corporations could adjust their policies to promote a healthier, lower-stress lifestyle among their employees.

Unfortunately, current methods for accurately detecting stress are intrusive and expensive. Such methods often cost thousands of dollars and typically involve attaching high-grade sensors and electrodes to an individual's body to continuously monitor physiological measures for changes that are indicative of stress. At best, this is impractical. Not only is the cost highly prohibitive, but it would be nearly impossible to go through a normal day with such equipment attached. At worst, it is impossible; insiders will not be wearing such equipment while perpetrating their misdeeds. What is needed is a cheap and accurate technique to detect stress without the use of such sensors.

The purpose of this thesis is to evaluate the potential of one such technique. Our objective is to use keystroke dynamics – the study of a user's typing rhythms – as a detection method for stress. The ubiquity of computing devices means that keystrokes are constantly being generated as a user goes about his or her daily business. Capturing these keystrokes, and their associated typing rhythm, is as easy as writing a simple piece of software to collect them. No specialized hardware is required; virtually every computing device has a physical or virtual keyboard built-in. Moreover, no explicit action is required from the user; a user can simply go about his intended task while the stress detection software runs in the background and s/he will not be inconvenienced in any way. If it is possible to detect stress via changes in typing rhythm, this technology could be rapidly and cheaply deployed while being invisible to end-users.

The work in this thesis is intended to act as a proof-of-concept, not to generate a finalized, deployable system. As such, the experimental protocols employed in this thesis are designed to rigorously evaluate the potential of keystroke dynamics as a stress detection method. The protocols are tightly controlled, with conditions that are likely unrealistic for real-world scenarios; however, the tight controls permit us to have high confidence in our outcomes by minimizing the deleterious impact of confounds and sources of noise. For the purpose of this thesis, we limit our scope to keystrokes on a standard computer keyboard. These keyboards are highly standardized and will be familiar to all our subjects. While this focus simplifies our experimental procedures, the experimental methodology and analytical techniques we use can be easily extended to other keyboards, including touch-screen and mobile devices.

It is also worth noting that the methodology we employ to detect stress using keystroke dynamics can be easily adapted to detecting other phenomena of interest. This could include the detection of other affective states, such as frustration, boredom, and excitement. The ability to detect a variety of affective states would be highly beneficial to the field of affective computing – which aims to create computing devices that dynamically respond to human emotion – or to develop more effective computer-based tutoring systems. Also of interest is the detection of physiological disorders (e.g., Carpal Tunnel Syndrome) or neurological conditions (e.g., Parkinson's disease, Alzheimer's disease, cognitive decline). Early-detection of these afflictions results in better outcomes, and the ability to monitor the progression (or regression) of the affliction on a daily basis would lead to more effective treatment plans. The primary difference between the detection of affective states and the detection of diseases is that affective states can be easily induced in laboratory subjects,

while it is neither practical nor ethical to impose afflictions on subjects. However, the fundamental idea remains the same: detecting phenomena of interest via changes in typing rhythms.

As a preview of our results, we found that each subject in our study has identifiable changes in typing that are associated with stress, though these changes seem to be highly individualized.

Chapter 2

Problem and Approach

Problem. The overarching problem addressed in this thesis is to determine whether or not stress is manifested as changes in an individual's typing rhythms, and whether such changes are universal across all typists, or across at least some typists. To achieve this, we break the task into several smaller pieces.

0. Induce stress in experiment participants.
1. Characterize how an individual subject's typing rhythms are affected by stress.
2. Identify universal markers for stress.
3. Identify groups of subjects that share common markers.

Our experiments were performed under tightly-controlled laboratory conditions. We induce stress with a computer-based multi-tasking user game, accompanied by social evaluation (e.g., a negative judgment about a participant's performance). We determine objectively how well the stressor worked by measuring independent physiological and psychological indicators of stress.

The first step in deploying keystroke dynamics to detect stress is to establish that we can reliably detect and characterize typing differences caused by stress. We start with the manifestations of stress in a single user's typing. A simple example would be that a 10% increase in typing speed characterizes the difference between when a user is under stress, as compared to being in a neutral setting. We use the term *feature* to refer to some aspect of the data that is changing; in this example, the feature is typing speed.

We use the term *marker* to refer to any easily-interpretable characterization that holds across groups of typists; we use the term to draw similarities to biological markers (e.g., the presence of the N-acetylgalactosamine antigen is a marker for Type A blood). In an ideal situation, we would find universal markers for characterizing the difference between neutral and stressed typing – markers that hold across an entire population (e.g., all users exhibit a 10% increase in typing speed when stressed).

If universal markers cannot be found, we will isolate smaller groups of users that share common markers. Stress may manifest in one group as a 10% increase in hold times (the duration a key is depressed), while it may manifest in another group as a 10% decrease in latency times (the duration between key presses). Assuming that no universal markers are present, we will identify all groups along with their markers.

Approach. Our approach will be an experimental one, as depicted in Figure 2.1. Data will be collected from subjects using a single-subject ABA (baseline-stress-baseline) design in which each
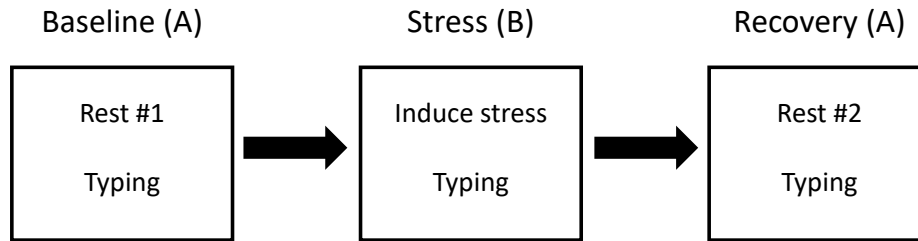
Figure 2.1: **Experiment overview.** A high-level overview of the three central phases of our experiment. Subjects undergo an initial rest period to place them in a neutral state, whereupon they provide a typing sample. Then, stress is induced in subjects through the use of a vetted stressor; subjects then provide a second typing sample. Finally, a second rest period returns the subject to a neutral state, where a third and final typing sample is provided.

subject acts as his/her own control. Subjects will provide typing samples – consisting of 80 repetitions of the phrase `great friends are good to have` – in each of three conditions: neutral/baseline, stress, and recovery/return to neutral/baseline. Concomitantly, physiological (e.g., blood pressure, heart-rate variability) and psychological (e.g., State-Trait Anxiety Index) data will be collected to provide independent and objective evidence that subjects are in the expected state when providing typing samples. Both neutral and stress states will be induced in the subject; the former via rest periods and the latter via a multitasking exercise coupled with social evaluation.

Extraordinary care was taken in the experimental approach to remove biases and to rule out potential confounds. In particular, care is taken to minimize unanticipated induction of affect in the subject. Confounds, such as the influence of practice, are accounted for in both the experimental methodology and in the analysis of the collected data. The entire study was tightly controlled.

Our analytical techniques comprise a mix of statistical and machine learning methods to identify changes in typing due to stress. These are employed in both a single-subject approach – comparing a subject's stressed typing sample against only his/her baseline typing samples – and also in a universal manner, with data from all subjects included. Clustering techniques are employed to identify groups of subjects that share common markers.

Anticipated outcome. If the experiments work, we would expect the following results:

0. Stressor. The stress paradigm is shown to have been effective if objective and independent measures of stress (e.g., blood pressure) are responsive to stress.
1. Typing. Users' typing rhythms show systematic changes due to induced stress; markers can be identified.
2. Universal markers. Markers that are consistent across users will be identified.
3. Subgroups. Markers that not universal, but are shared by subgroups of the experiment population, will be identified.

Chapter 3

Related Work

We divide related work into three major categories: (1) research on the use of keystroke dynamics for tasks other than affect or valence detection, (2) research on the use of keystroke dynamics for affect or valence detection, and (3) research on stress, focusing on work that describes the effect of stress on fine motor control, as would pertain to typing. Affect refers to the emotion(s) felt by the subject (e.g., happiness, sadness, frustration). Valence refers to the intrinsic attractiveness (positive valence) or repulsiveness (negative valence) of a stimulus.

3.1   Traditional keystroke dynamics

Although we are using keystroke dynamics for stress detection, much of the keystroke literature is focused on discriminating amongst different individuals, usually for computer-security purposes, briefly reviewed here. Keystroke dynamics relies principally on two features: hold times (the length of time a key is held down) and latency times (the time between one key being released and the next key being depressed). Other features, such as combinations and ratios, can be derived from these.

Keystroke dynamics traces its roots to Bryan and Harter (1897), who demonstrated that telegraph operators could be discriminated based on their keying rhythm. The first work using a computer keyboard was conducted by Gaines et al. (1980), who discovered that the timings of just 5 digrams were sufficient to perfectly discriminate amongst 7 subjects. Most research in keystroke dynamics aims to discriminate between users under a variety of circumstances. Many of the recent developments have focused on short, fixed texts (e.g., passwords), which are quick to collect and easy to analyze (Peacock et al., 2004). Some results have been promising: Obaidat (1995) claimed to perfectly discriminate between 15 subjects using a neural network, and Yu and Cho (2003) obtained a false reject rate of 0.814% with a false accept rate of 0% when discriminating between 21 subjects using a Support Vector Machine. It is worth noting that when comparing keystroke algorithms on a fair basis (e.g., using the same data set, which researchers tend not to do), simple techniques turn out to work better than the more complicated ones; for example, Killourhy and Maxion (2009) showed that a classifier based on scaled Manhattan distance was the best performer out of 14 classical methods.

Recent work has focused on long-text (e.g., paragraphs). Gunetti and Picardi (2012) obtained an error rate of 0.5% when discriminating amongst 311 users who composed messages of roughly 65 characters, while Samura and Nishimura (2009) obtained 100% accuracy in discriminating between 52 subjects who composed Japanese free text. A recent survey of the field was conducted

by Teh et al. (2013).

3.2   Keystroke dynamics for non-stress affect or valence detection

There has been a handful of works focusing on affect or valence detection using keystroke dynamics, where stress is not the affect of interest. Unfortunately, all of these works have one or more critical shortcomings, such as conducting an uncontrolled experiment, the use of non-vetted affect induction techniques, lack of an independent measurement of affect, and/or lack of proper statistical testing. Due to these shortcomings, the resulting work is neither rigorous nor valid; in many cases, the work is not even reproducible. Nevertheless, we briefly review it because it is the best we can find.

To our knowledge, ten research groups have attempted to discriminate between different affective or valence states. Alhothali (2011) differentiated among five affective states: delighted, neutral, confused, bored, and frustrated. Epp et al. (2011) differentiated among 15 affective states, including anger, excitement, nervousness, sadness, and tiredness. Fairhurst et al. (2015) differentiated between happy and non-happy states. Kolakowska (2015) differentiated among 7 affective states: fear, anger, sadness, happiness, disgust, boredom, and surprise. Lv et al. (2008) differentiated among six affective states: neutral, anger, fear, happiness, sadness, and surprise. Nahin et al. (2014) differentiated among seven affects: joy, fear, anger, sadness, disgust, shame, and guilt. Shikder et al. (2017) differentiated among ten affective states, including amusement, happiness, inspiration, sympathy, and disgust. Tsihrintzis et al. (2008) differentiated among six affects: neutral, surprise, anger, happiness, sadness, and disgust. Khanna and Sasikumar (2010) differentiated among positive, negative, and neutral valence. Finally, Zimmermann et al. (2003) differentiated among positive, neutral, and negative valence and between low and high arousal states. Reported accuracies ranged between 57% (Tsihrintzis et al., 2008) to 95.6% (Lv et al., 2008).

None of the groups used a vetted affect induction technique along with vetted stimulus materials. Techniques employed ranged from having subjects use an automated computer tutor (Alhothali, 2011), read text passages (Lv et al., 2008), or watch movies (Zimmermann et al., 2003). Eight groups either did not independently confirm affect induction or relied on self-reporting. Alhothali (2011) asked judges to subjectively determine affective state based on videos of the subjects. Zimmermann et al. (2003) mentioned using physiological parameters as an objective measurement of affect; however, no results were presented in their paper.

Only one group (Zimmermann et al., 2003) described a controlled experiment. The other groups either stated that their experiment was uncontrolled, or reported so little methodological detail that it is impossible to ascertain whether a controlled experiment was conducted. Details such as the instructions to subjects, stimulus items, and descriptions of data collection are generally omitted. Without such details, it is impossible to judge the merits of the research conducted. In the proposed work, of course, we will undertake to correct the shortcomings noted here.

3.3   Keystroke dynamics for stress detection

A brief survey of stress detection through keystroke and mouse dynamics has been performed by Kolakowska (2013). We focus here on the papers that primarily used keystroke dynamics and we provide a brief discussion on the validity of each work.

Andren and Funk (2005) performed a proof-of-concept experiment with 4 subjects to determine whether stress could be detected via keystroke dynamics. Subjects provided typing data in both neutral and stressed conditions. No details were provided about the stressor, nor were details provided about any objective measurement of stress. Based on reported results, it cannot be determined whether or not the stressor was vetted or whether an objective measurement of stress was used. Classification accuracies ranged from 20% to 100%; the wide range of accuracies could be explained by the methodological shortcomings in the study.

Bando et al. (2015) performed a study with 18 subjects to determine whether there were correlations between two physiological measures of stress – heart rate variability and respiration – and keystroke latency times. Stress was induced by exposing subjects to white noise, which is not a vetted stressor. They concluded that such correlations do exist, though it is not clear if these correlations are significant.

Gunawardhane et al. (2013) performed a study on 20 students to determine whether typing, as measured by bi-graph and tri-graph duration times, produced in a stressed state was significantly different from that produced in a neutral state. Non-stressed typing data were collected from the students when they were not preparing for exams; stressed typing data were collected during an exam period. Students were exposed to a mental arithmetic test prior to the stressed data collection to add additional stress. An attempt was made to confirm stress induction via a self-developed questionnaire administered to the subjects and also using a proprietary "stress monitor". It is unclear how well either instrument works. The results presented are in terms of p-values – there are significant differences between stressed and neutral typing on several bi-graphs and tri-graphs (e.g., 'a-n' and 't-h-e'). Unfortunately, there appears to have been no correction for multiple testing even though many tests were performed (at least 50); this casts doubt upon the presented results.

Kolakowska (2016) performed a study on 16 subjects to determine whether stress has significant effects on digraph and trigraph latency times. Typing was collected from subjects as they completed two coding tasks, one without time pressure followed by one with time pressure. The presence of the time pressure was postulated to induce stress in the subjects, though no objective measures of stress were used. The study concluded that digraph and trigraph latencies were statistically significantly affected by the presence of stress.

Lim, Ayesh, and Stacey have published a series of five papers regarding the correlation between perceived stress and keystroke/mouse behaviors (Lim et al., 2014a,b,d,c, 2016). There appear to be two underlying experiments that were conducted. The first study had 60 subjects asked subjects to answer mental arithmetic questions with and without a time constraint. The second study also had 60 subjects and involved a variety of transcription tasks, again with and without a time constraint. Placing a time constraint on the subjects was the only method of inducing stress; it is unclear whether time constraints alone are sufficient to induce stress. No objective measure of stress was obtained in either study, only a subjective response from each subject. The studies established statistically significant correlation between perceived subject stress and changes in average keystroke latency times.

Vizer et al. (2009) performed an exploratory study with 24 subjects to determine whether stress could be detected via a combination of keystroke dynamics and linguistic analysis. They attempted to distinguish between a neutral condition and a physical stress condition and between a neutral condition and a cognitive stress condition; we focus on the cognitive stress condition since that is most similar to the proposed work. An attempt was made to induce cognitive stress via a mental multiplication task and a three-back number recall task. Despite the claims made in the paper,

it is important to note that neither task is considered a vetted technique for inducing cognitive stress; rather, these tasks are vetted tests for working memory. As such, they are less appropriate as stress-induction methods. We also note that there was no objective measure of stress in the study. Several different classifiers were employed; the best result was 75% accuracy when using a *k*-nearest neighbor algorithm. The analytical focus in the Vizer work was on linguistic features, not typing features, so the extent to which keystroke analysis played a central role was unconfirmed.

3.4   Stress and fine motor control

There are several potential pathways between stress and fine motor control. The first, and most well-understood, is through the hypothalamic-pituitary-adrenal (HPA) axis. When under stress, the hypothalamus releases corticotropin-releasing hormone (CRH), stimulating the production of adrenocorticotropic hormone (ACTH), which in turn increases the production of cortisol, epinephrine (also known as adrenaline) and norepinephrine. The presence of epinephrine causes an increase in the contractile force of the skeletal muscles. This can directly influence typing by causing actions to be undertaken with increased force and also indirectly by magnifying physiologic tremors. It is worth noting that adrenaline and noradrenaline are secreted via direct innervation between the hypothalamus and the adrenal medulla (hence the fast acting response) - via the Sympathetic-Adrenal-Medullary (SAM) axis, rather than the (relatively slower) HPA axis with the end result of cortisol.

The remaining pathways operate through the neurotransmitter, dopamine. During stress, the rate of dopamine loss and replacement is increased, altering the levels of dopamine in the brain. Dopamine plays a vital role in regulating motor control and motor-path-planning processes. It is believed that dopamine influences which motor actions are taken, though the precise pathways are unclear. Obviously, any change in the motor actions will be reflected in typing.

Finally, dopamine is known to play a role in tremor, most notably in the case of Parkinson's Disease (PD). PD is marked by the death of dopamine-generating cells in the substantia nigra, which is thought to cause the rest tremors that are most commonly associated with the disease. The precise mechanisms underlying dopaminergic tremors are unclear, but a deficiency of dopamine, caused by stress, may cause or exacerbate tremors.

We are not the first researchers to suggest a link between stress and fine motor control. There is a sizable literature on techniques for reducing the effects of stage fright (a form of stress) on the performing capabilities of musicians (Lehrer, 1987). Similar research has been conducted on the performance of skilled shooters (Lakie, 2010). In both musicians and shooters, stress is believed to increase the magnitude of tremors and to affect the tension levels of the muscles in the arm and hand. Such changes have deleterious effects on performance.

Chapter 4

Data

The data collected as part of this thesis work fall into three separate categories: 1) independent-validation data, 2) typing data, and 3) supporting data.

Independent-validation data include physiological measures and self-reported, short-term psychological measures that provide evidence that the typing data were actually collected in the expected affective states – either baseline neutral or stressed.

Typing data are essential in addressing the problems posed in this thesis work. It is in these typing data that we expect to see manifestations of stress that constitute the central focus in this thesis.

Supporting data consist primarily of responses to questionnaires, video and photographic evidence, and physical measurements of our subjects, that may contain explanatory value for phenomena discovered in the course of analyzing the typing or supporting data. Such data may explain, for example, an anomaly in the collected data, why affect induction may have been unsuccessful for a particular subject, or why particular subjects show common manifestations of stress in their typing.

We start with a high-level description of the data before diving into details of how these data were collected and pre-processed prior to analysis. For many of these data types, the raw data (i.e., data in its initially collected form) are not immediately suitable for analysis. Pre-processing must be performed to convert these data into a form more directly suited for analysis. Preprocessing details are relegated to the appendix.

As a reminder, all data were collected in tightly-controlled experiments. Protocols and forms are in the appendix; apparatus and detailed descriptions of experimental methods are in Chapter 5.

4.1 Independent-validation data

Independent-validation data consist of physiological and self-reported, short-term psychological measures collected during the course of the experiment. Physiological data consist of blood pressure readings, an electrocardiogram, and respiration data. The psychological measures consist of subject responses to two questionnaires: 1) the short-form State-Trait Anxiety Inventory (STAI) and the 2) NASA Taskload Work Index (NASA-TLX).

Blood pressure. Each blood pressure reading consists of a quadruplet of systolic blood pressure, diastolic blood pressure, pulse rate, and mean arterial pressure. An initial reading is taken prior to the start of the experiment to ensure that the subject conforms to the inclusion criteria. During the experiment, blood pressure readings are taken at 5-minute intervals during the two rest

periods and during the stress induction period. It is expected that all four of these measures will be elevated when a subject is stressed, as compared to baseline.

Electrocardiogram (ECG). ECG data are collected continuously during the course of the experiment using a modified lead-II electrode placement. The data are sampled at a rate of 10,000 samples per second. These data are then converted to the measures of heart-rate variability, which show marked changes between baseline neutral and stressed states. It is expected that the median R-R interval – which can be thought of as the time between consecutive heart beats – will be lower when a subject is stressed. Similarly, the SDRR – a measure of consistency in heartbeat timings – is also expected to be lower when a subject is stressed.

Respiration. Respiration data are also collected continuously during the course of the experiment using a respiration belt attached to the subject's midsection. The data are sampled at a rate of 10,000 samples per second from which a breaths/min value is extracted. It is expected that a subject's respiration rate will increase when stressed.[1]

Short-form State-Trait Anxiety Inventory (STAI). On six different occasions in the experiment – after each rest period, after each typing sample, and after the multi-tasking exercise – the subject provides responses to each of the 6 questions on the short-form STAI. Questions touch on the immediate psychological state of the subject, whether the subject is content, worried, upset, etc. Responses are in the form of vertical marks on a visual analogous scale, which are then converted to numeric measures by measuring distance from the left end of the line. It is expected that measures associated with stress (e.g., worry) will be higher when a subject is stressed, while measures associated with relaxation (e.g., contentedness) will be lower when a subject is stressed.

NASA Taskload Work Index (NASA-TLX). The subject responds to the NASA-TLX on the same six occasions when the short-form STAI is administered. The NASA-TLX also contains 6 questions, touching on the subject's experienced difficulty in completing a task. The NASA-TLX is also a visual analogous scale; as with the short-form STAI, vertical marks are also converted to numeric measures by measuring distance from the left end of the line. It is expected that measures of workload challenge (e.g., effort required) will be higher following the multi-tasking exercise while measures of performance will be lower following the exercise.

4.2 Typing data

Each subject in the experiment provides typing data on four different occasions, referred to as *sessions*. In each session, the subject is asked to provide correctly-typed repetitions of the phrase `great friends are good to have`. The choice of this particular phrase was quite involved; however, we defer discussion of how the phrase was selected to Section 5.4. Briefly, the phrase was selected to be easy to type so as to minimize the amount of variation in a subject's typing. Typing data are collected using custom hardware and software, as discussed in the appendix in Section A.1.

The typing data are broken down into keyup and keydown events, denoting the release or depression of a key, respectively. All key events – whether they correspond to correctly or incorrectly

---

[1]Though we collected respiration data, we did not end up using it in any of our analysis. There are two primary reasons. First, we directly instruct subjects to control their breathing in one of the rest periods in the experiment. Consequently, respiration would no longer be a fully independent measurement of affect. Second, the respiration data were extremely noisy, with numerous movement artifacts. A significant investment of resources would have been required to clean the respiration data. Given the high cost of using the respiration data and the low benefit, we opted to simply omit it from our analyses.

typed characters – are recorded by our software. Each key event is stored as a triplet containing 1) the ASCII code of the key, 2) whether this was a keyup or keydown event, and 3) the timestamp of the event. Each key event is recorded by two, paired logs. The first log utilizes a human-readable XML format with low-resolution timestamps. The second log is a coded, non-human-readable format which contains high-resolution timestamps. These logs are ultimately merged into a single log that is both human-readable and possesses the high-resolution timestamps. The details of this process can be found in Appendix A.2.

As there are four typing sessions, four sets of logs are generated, one for each session. Each log will contain either 40 (familiarization session) or 80 (baseline, stress, and recovery) correctly-typed repetitions of the phrase "great friends are good to have". Additional key events, corresponding to typographical errors, are also present in these logs.

4.3   Supporting data

Supporting data fall into six rough categories: 1) demographic, 2) psychological, 3) video, 4) photographic, 5) physical, and 6) observational notes. The value of this data is in explaining phenomena discovered in the course of analyzing the typing or independent-validation data. For example, changes in subjects' typing rhythms due to stress may be dependent upon the handedness of the subject; keys struck by the dominant hand of the subject may be less affected by stress. Anomalies in typing or physiological measures could also be explained by examining video data; a sneeze or cough could be responsible for an unusually lengthy typing repetition or for an unusual pattern in respiration data. These data are largely collected under the guiding philosophy of maximizing the research value of each subject run. While there may not be any specific plan for analyzing these data, they are collected so that we can reference them if they occasion to be of use.

Demographic data. Demographic data consist of responses to a questionnaire administered to each subject. Twenty-two questions were asked regarding topics with possible influence on a subject's typing behavior (e.g., gender, handedness, commonly-used keyboards). Some questions merit a free-form response; for others, subjects circle an answer from a pre-set list of questions. The questionnaire is shown in the appendix.

Psychological data. Psychological data consist of responses to two separate questionnaires: 1) the long-form State-Trait Anxiety Inventory and 2) the Perceived Stress Scale. As the name suggests, the former is related to the short-form STAI. Where the short-form STAI focuses on a subject's immediate psychological state, the long-form STAI focuses on how a subject's state in general. The Perceived Stress Scale, as the name suggests, covers a subject's perceived stress levels in day-to-day life. Responses to both questionnaires are collected using a modified Likert scale.

Video data. Four different streams of video, from four different cameras, are collected as part of the experiment. Three of the cameras are positioned to the left, to the right, and above the keyboard and are intended to capture the subject's typing behavior from different angles. The fourth camera is focused on the subject's face and is intended to aid in determining the status of the subject at any point in the experiment. The chief purpose of the video data is to enable us to revisit a specific point in time where an anomaly in the typing or supportive data occurred, so that we can search for any explanatory cause.

Photographic data. Two types of photographs are taken during the experiment. The first are photographs of each of the subjects' hands, with a calibrated grid in the background. This enables us to precisely determine the hand geometry of a subject, with the most notable measures being

hand sizes and finger lengths. The second type of photographs are taken using a still camera placed to the side of the subject. These photos capture the natural sitting posture of a subject as s/he types. Both types of photographs could potentially provide evidence that explains why some subjects had typing changes not present in others.

Physical data. We measure the height and weight of our subjects as part of the experimental protocol; from these we can calculate BMI (body mass index). This information can be used to more finely calibrate the blood pressure and electrocardiogram data.

Observational notes. Our experimental protocol mandates that the experimenter makes a series of notes regarding the subject's general and typing behavior. General behavior notes include observations on the general demeanor of the subject and any information gleaned during casual conversation that may be of explanatory value (e.g., a subject is generally agreeable vs. unusually antagonistic). Typing behavior notes may include the general force with which a subject types or unusual fingerings of the keys (e.g., using the right index finger to hit the Return key). Any unusual occurrences in the experiment will also be logged in the observational notes; these may include things like an error in the blood pressure reading due to a subject flexing his/her arm repeatedly during a reading.

Chapter 5

Experimental Methods

We turn our attention now to the experimental methods used in the course of this thesis. As a prelude, we start with a brief, high-level overview of the experiment; the intent is to provide context for the materials in the remainder of this chapter. The meat of the chapter consists of a description of the data collected as part of the experiment, a discussion of the apparatus and instrumentation used in the collection of these data, the methods by which the stimulus text was chosen for the experiment, the experimental design employed, the precise details of the protocol used for each subject in the experiment, and the instructions provided to our subjects.

5.1   Guiding philosophy

In designing this study, we were primarily concerned with its internal experimental validity. That is, we want our conclusions to be as free from potential confounds, sources of noise, and bias as is possible. We are engaged, effectively, in vetting the promise of a new technology: the use of keystroke dynamics as a detection mechanism for stress. Our goal is to assess, as accurately as possible within our abilities, how well this technology works. If confounding factors or significant noise were to be present, it would significantly diminish the scientific value of our study. A conclusion that keystroke dynamics is or is not capable of detecting stress is rendered meaningless if we cannot be confident that the reasons for success or failure are directly related to the actual efficacy of keystroke dynamics in stress detection. Conducting this experiment is a significant expenditure in time, resources, and personnel; we want to be really sure that we get this right, as we only have a single chance.

Our primary concerns in designing the study were:

Minimizing confounds. A confound is any unaccounted-for variable that could have a significant bearing on the conclusions of the experiment. In the context of the present experiment, confounds would include variables that influence a subject's response to stress. Examples of such variables would include caffeine, alcohol, or drug consumption; mental illnesses such as anxiety disorders; and time of day when an experiment is conducted (as the main stress hormone, cortisol, has a natural diurnal cycle that peaks in the morning, known as the cortisol awakening response).

Minimizing bias. Bias refers to any process that may prejudice a particular outcome in our experiment. For example, suppose that we were to use an inadequate stressor that did not actually induce stress in any of our subjects. We would then find little difference between the collected neutral and stress typing from our subjects, leading us to conclude that stress cannot be detected through keystrokes, when stress may very well manifest in keystrokes. This inadequate stressor would have

biased us towards the incorrect conclusion that stress cannot be detected through keystrokes. In our study, we employ a validated stressor that has been demonstrated to work, thus avoiding this particular bias.

Maximize data collection. This experiment was designed with the knowledge that it would require significant time investment. Our initial intent was to gather data from approximately 130-140 subjects, with the understanding that we could only run a single subject per day and only on weekdays. Accordingly, even with perfect scheduling, data collection would have to last over 6 months; taking into account cancellations, dropouts, no-shows, recruiting non-responses, holidays, and blank calendar spots, data collection could easily last a full year. Given this mandatory investment of time and resources, we wanted to ensure that we collected the maximum amount of data possible for each subject. In several cases, this has resulted in the collection of data that are not directly analyzed in this thesis; however, the data is available for future researchers to analyze, if they believe it has value.

Consistency between subjects. This experiment was, by far, the most complicated one that we have attempted to conduct in our lab. Our prior experiments generally consisted of a subject providing a single typing sample, with each sample taking less than 20 minutes. By contrast, this experiment includes four typing samples, along with stress induction and the recording of numerous physiological measures; a single subject for this experiment takes between 2 to 3 hours to run. Given the relative complexity and length of each experimental session, we took care to design the experiment in a manner that enabled us to provide a consistent experience to each subject. All subjects were run under identical lab and environmental conditions.

5.2 Experimental Overview

At a high level, this experiment is relatively simple. A subject provides typing data on four occasions: 1) during a familiarization period, 2) in a baseline neutral state, 3) in a stressed state, and 4) in a recovery neutral state. To find the desired markers for stress, the data from the neutral states and the stressed state are compared.

As always, the devil is in the details. Of critical import in this experiment is the affective state of the subject when providing these typing samples. The conclusions we draw will be invalid if we cannot be sure that a subject is actually in a baseline neutral state or in a stressed state, as appropriate, when providing the typing samples. Two rest periods, each immediately preceding a neutral typing sample, are employed to bring subjects to a neutral state. In the first rest period, subjects watch a relaxing movie while performing a simple task. In the second, subjects are asked to focus and control their breathing. Stress is induced in our subjects through the use of a timed multi-tasking exercise in conjunction with social evaluation (negative social judgment). A variety of physiological measures (e.g., blood pressure and an ECG) and self-reported psychological measures (standard stress and workload surveys) are collected to independently verify that the rest periods and stressor have had their desired effects.

The success of the experiment hinges on the particulars of its execution. Within this chapter, we offer considerable detail regarding the conduct of the experiment. While such detail may be common in other disciplines, we acknowledge that it is not common within computer science and machine learning. We offer these details because they can best reveal potential biases, confounds, and threats to experimental validity. Moreover, we do so in the interest of other investigators being able to replicate our work, to judge the correctness of our work, and to take up where we have
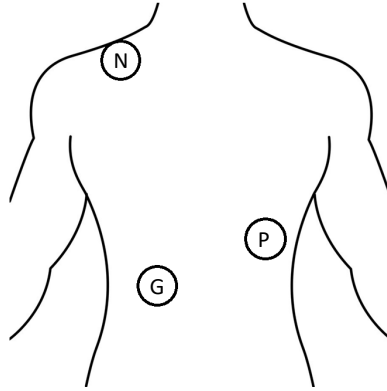
Figure 5.1: **Modified Lead-II placement.** Three electrodes are used in a modified lead-II placement. The negative electrode (N) is placed on the bony part of the right shoulder, the positive electrode (P) is placed on the lower-most left rib, and the ground electrode (G) is placed on the right-side of the torso, slightly vertically lower than the positive electrode. The modified lead-II placement maximizes the prominence of the R-wave in an ECG.

left off. We do not expect a reader to read the entirety of this chapter, unless s/he is curious or interested in the particulars of the methodology supporting the experiment.

The apparatus and instrumentation, both for creating and measuring the affective state of the subject, are detailed in Section 5.3. A brief overview of the choice of typing stimulus (i.e., Why do subjects type the phrase `great friends are good to have`?) is provided in Section 5.4. The inclusion and exclusion criteria, as well as the procedures for subject recruitment are detailed in Section 5.6. The particulars of our protocol are discussed in great detail in Section 5.8. Finally, instructions provided to our subjects are detailed in Section 5.9.

5.3   Apparatus and instrumentation

The apparatus and instrumentation in this experiment serve two primary purposes. First, the apparatus and instrumentation is used to collect the numerous types of data in the experiment. Second, they are used to induce the desired affective states – either neutral or stressed – in our subjects. We start with a breakdown of the employed equipment by the data types they are used to collect. This is followed by a discussion of the equipment used to encourage a relaxed state or to induce a stressed state in our subjects.

5.3.1   Independent validation data

Independent validation data are separated into two categories: 1) physiological data and 2) psychological data. Physiological data provide an objective measure of the affective state of the subject while psychological data provide a subjective, self-reported measure of a subject's affective state.

Physiological data

Three different types of physiological data are collected in our experiment: 1) blood pressure, 2) electrocardiogram, and 3) respiration. Blood pressure is collected via a medical-grade blood-pressure monitor (model number: GE Healthcare Carescape V100 Vital Signs Monitor); this is a standard clinical blood-pressure machine found in professional laboratories and hospitals. Two

blood-pressure cuffs are employed for the sake of the experiment – Adult Medium and Adult Large. The cuff size is chosen based on the girth of the subject's arm; most subjects are adequately served by the Adult Medium size (less than 2% of subjects required the larger cuff). Prior to application of the blood pressure cuff, the subject is asked which hand they use for the mouse. The cuff is placed on the opposite arm; this is typically the non-dominant arm, but this is not always the case. Blood pressure readings tend to fail if the arm is in motion; thus, obtaining blood pressure readings during the stressor task requires us to attach the cuff to the non-dominant arm. Blood pressure readings are triggered manually by the experimenter at 5-minute intervals during the rest periods and stressor task. We had considered automatically triggering the readings through the LabChart software (discussed below), but we were thwarted by insufficient and inconsistent documentation of the blood-pressure machine's inner workings.

The electrocardiogram and respiration data are both channeled through the LabChart Pro 8 software, purchased from AD Instruments (AD Instruments, 2014). Electrodes for the electro-cardiogram are placed according to a modified lead-II placement, as indicated in Figure 5.1; this placement was chosen as it maximizes the prominence of the R-wave (Stern et al., 2001). The neg-ative electrode is placed on the bony part of the right clavicle (collar bone), the positive electrode is placed on the lower-most left rib, and the ground electrode is placed on the right side of the torso 2 inches below the positive electrode. Subjects are asked to assist the experimenter in locating the relevant positions for the negative and positive electrodes. The signal passes through a Dual Bio-Amp into a PowerLab 16/25, both manufactured by AD Instruments. Respiration data are col-lected using a Polar Respiration Belt (Respiration Belt, 2014), routed through the same PowerLab 16/25. The respiration belt is attached in a snug fashion around the subject's waist, right above the belly button. Care is taken to ensure that the fit is snug during regular breathing; the subject is asked to ensure that taking a deep breath results in slight discomfort, as this confirms the belt is sufficiently tight. Both the ECG signal and respiration signal are sampled at a rate of 10,000 samples per second.

We chose to use the LabChart software and hardware because of its open source nature. We had originally considered competing products from other vendors as well. However, the competing products used proprietary file formats. Since we wanted to be able to freely access and analyze the data using whatever methods we pleased, we opted to use LabChart since it uses an open-access file format (CSV).

Despite our previous mention of cortisol (the stress hormone) and the cortisol awakening re-sponse, we did not collect cortisol samples. The two primary measures for cortisol are plasma and salivary cortisol. Collection of plasma cortisol requires blood to be drawn from the subject. Such a setup would be expensive and impractical; moreover, we lack personnel with the appropriate training to draw blood and lack proper storage facilities for drawn blood. Collection of salivary cortisol was deemed to be unwise for three reasons. First, it is a lagging indicator of stress, so we were concerned that it would not adequately reflect the stress state of our subjects. Second, we already had independent measures of stress by monitoring ECG and blood pressure. We felt that also collecting salivary cortisol would have added unnecessary complications to an already com-plex experimental protocol. Finally, we did not have pre-existing refrigeration storage for salivary cortisol and obtaining the required space for such storage would have been extremely difficult.

#	Statement
1	I feel calm
2	I feel tense
3	I am upset
4	I feel relaxed
5	I feel content
6	I feel worried

Table 5.1: **Short-form STAI.** Subjects are asked to rate their agreement to 6 statements on a visual analogue scale of "not at all" to "very much" at this particular moment in time.

#	Statement
1	Mental demand
2	Physical demand
3	Temporal demand
4	Effort
5	Performance
6	Frustration

Table 5.2: **NASA-TLX.** Subjects are asked to rate their experience of a just-completed task on 6 axes by marking a visual analogue scale of "low" to "high".

Psychological data

Two types of psychological data are collected: (1) general and (2) task-dependent. Each type of data consists of responses to two different questionnaires.

General psychological data consist of responses to the 1) long-form State Trait Anxiety Form Y-2 (STAI-Y) and the 2) Perceived Stress Scale-10 (PSS-10). The STAI Form Y-2 (Spielberger et al., 1983) is designed to capture a subject's general anxiety in day-to-day life (e.g., "I am satisfied with myself"; "I feel that difficulties are piling up so that I cannot overcome them").

The PSS-10 (Cohen et al., 1983) is designed to capture a subject's perception of the degree to which different types of stressors are present in his life (e.g., "How often have you felt that things were going your way?"; "How often have you been angered because of things that happened that were outside of your control?"). Both forms are administered once at the start of the study. The general psychological data permit us to examine whether the magnitude of manifestations of stress correlate with a subject's general stress level.

Task-dependent psychological data consist of responses to the 1) short-form State Trait Anxiety Inventory (STAI) (Marteau and Bekker, 1992) and the 2) NASA Taskload Work Index (NASA-TLX) (Hart and Staveland, 1988). The short-form STAI consists of 6 questions, depicted in Table 5.1, asking a subject to self-report levels of stress and anxiety at the current moment on a scale of "not at all" to "very much"?. The NASA-TLX, also 6 questions (Table 5.2), asks a subject to self-report the level of workload felt in the previously completed task (from "Low" to "High"). Both forms are visual-analog scales (VAS), where the response to each question is a vertical mark on a 100mm line; the ends of each line represent the extremes (e.g., "not at all" and "very much"). These self-reports are essential to the study, as they permit us to evaluate the effect of the rest

periods and stressor task. It is expected that subjects will feel low levels of stress and anxiety, along with low workload, after rest periods. It is expected that subjects will feel higher levels of stress and anxiety, along with higher workload, after the stressor task. The task-dependent psychological data complement the physiological data that are collected. The former allow us to confirm a subject's subjective feeling about his stress state; the latter allows us to objectively confirm a subject's stress state.

It is worth noting that both the short-form STAI and the NASA-TLX are quick to complete. It would be reasonable for a subject to complete both forms in under 30 seconds. This is particularly useful in our experiment, as we are concerned with a subject's stress levels diminishing while they complete the forms.

5.3.2 Typing data

Typing data are collected using custom software and a modified external keyboard. The apparatus used in this experiment was designed with two major goals in mind: 1) to allow maximal control over the presentation of the stimulus to the subject and 2) to collect highly accurate timestamps. The custom software, called MTP, is responsible for displaying instructions and displaying stimulus prompts to the subjects. In this particular experiment, stimulus prompts consist entirely of repetitions of the phrase 'great friends are good to have'. For each repetition of the typed phrase, the subject is presented with a blank text box in which s/he must correctly type the phrase, followed by the Enter key. All characters in the phrase must be typed correctly, in sequence. Any typographical errors made by the subject cause the text box to become momentarily grayed out, then cleared; the subject must then re-type the phrase from the beginning. The MTP software runs on a laptop running Windows XP; the machine has been stripped of as many processes as possible, so as to reduce the effect of system load on the collected data. The machine is also disconnected from the network, to reduce the influence of network interrupts on system performance.

The hardware employed in this experiment consists of a modified Apple USB-keyboard (model number: M9034LL/A) with a standard QWERTY layout. The keyboard was modified by removing the standard keyboard controller and rerouting the output through a custom external timer (colloquially referred to as the "Gizmo"). The Gizmo provides timestamps with an accuracy of 200 microseconds. This accuracy was confirmed by using a function generator to simulate keystrokes at a fixed interval.

Whenever the subject presses or releases a key – regardless of whether the key is correct or not – the key is recorded by two different logging processes. The first is contained within the MTP software and the second is connected to the Gizmo external timer. The MTP log contains coarser timestamps, but the log is formatted in a manner that is human-readable and contains meta-information, such as what stimulus was presented to the subject. The Gizmo log contains the exact, 200-microsecond-accuracy timestamps, but is formatted in an encoded format; it does not contain any of the meta-information present in the MTP log.

This dual-logging procedure permits a reconciliation of the timestamps at a later time, detailed in Appendix A.2, to produce a single log that contains both the highly-accurate timestamps and the meta-information. Note that it is far easier to perform this reconciliation at a later time, and not in real-time, since the two logging processes record slightly different things; the MTP log contains only keystrokes sent to the MTP application, while the Gizmo log contains all keystrokes struck, regardless of the target application. While there is considerable overlap between these two logs, keystrokes associated with starting or terminating the MTP application are generally

recorded only in the Gizmo log and not the MTP log. Additionally, if MTP ever loses focus as the active application – an admittedly rare occurrence in the tightly-controlled laboratory setting of this experiment – keystrokes will be recorded only in the Gizmo log and not the MTP log. This provides some redundancy in our data collection and permits error correction.

5.3.3   Supporting data

Supporting data are separated into three categories: 1) demographic data, 2) physical data, and 3) video data. These data are gathered because they may hold explanatory value for observed anomalies or trends in the typing or supporting data.

Demographic data

Subjects in the study are asked to fill out a 22-question demographic questionnaire. This demographic questionnaire was derived from previous questionnaires that have been used in other studies our lab has conducted. Questions are designed to elicit information about physical or behavioral factors that may have explanatory post-analysis capability. Of particular interest is whether clusters of subjects with similar manifestations of stress are related by one or more factors. For example, are all subjects with a large manifestation of stress left-handed? Identified factors may also allow us to explain anomalies in our subject pool. A subject with an unusual manifestation of stress may also have a physical injury; the injury could possibly explain the odd manifestation.

The first set of questions is largely generic, focusing on attributes like age, gender, and education. These factors could potentially explain differences in manifestations of stress. Older subjects and more educated subjects may be less prone to stressors as they have been in more stressful situations. Biological differences between men and women may result in differing responses to stress.

The second set of questions focuses on typing behavior. Subjects are asked questions such as how they learned to type and how much time they spend typing. Subjects who learned to type at an early age and those who spend significant time typing may exhibit smaller manifestations of stress; typing for such subjects would be a highly practiced activity that may be less affected by stressors.

The third set of questions focuses on physical traits that may influence typing behavior. For example, subjects are asked whether they have long fingernails or wear jewelry that may influence typing. Subjects with long fingernails generally have substantially different typing patterns from those with short fingernails (e.g., it is generally highly unpleasant for a subject with long fingernails to strike the key with the tip of the nail). Subjects wearing cumbersome jewelry also generally have altered typing patterns; we ask subjects to remove such jewelry. These factors could be potential confounds in our experiment; by asking these questions, we can account for the influence of these factors when we perform our analyses. The questionnaire also asks whether subjects suffer from temporary or permanent conditions that may influence typing. Temporary conditions might include a sprained finger or swollen joints; permanent conditions could range from shortened or missing fingers to arthritis. Clearly, such conditions will have a significant impact on a subject's typing overall; the manifestation of stress in that subject's typing may likewise be affected.

The final set of questions concerns the subject's experience with various styles and layouts of keyboards. In this experiment, subjects are asked to produce their typing samples on a standard desktop QWERTY keyboard. While most subjects use such a keyboard (or the laptop equivalent) on a regular basis, some subjects used other styles and layouts in their day-to-day activities. Sub-
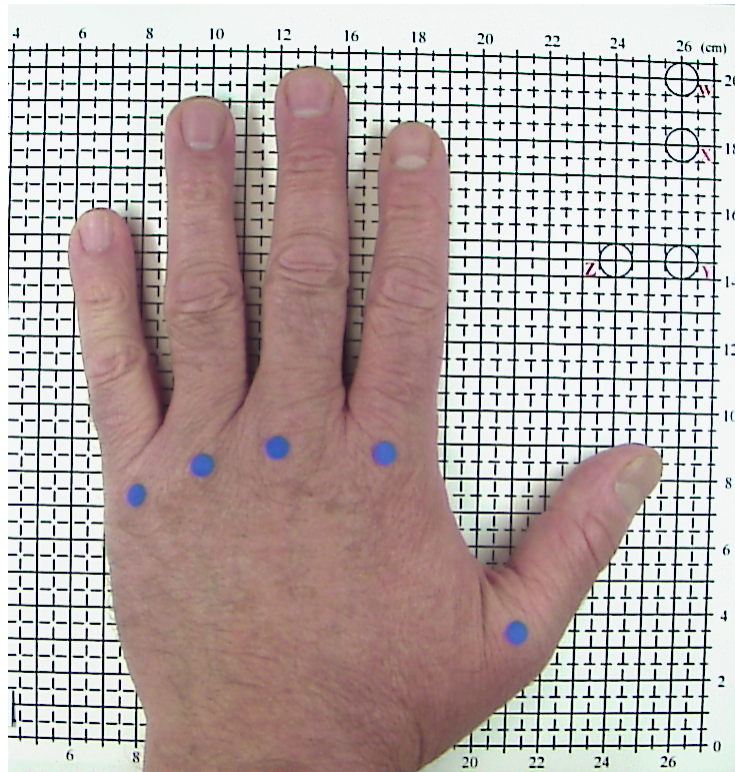
Figure 5.2: **Dot placement for hand photos.** The positions of the dots can be used to accurately extract the dimensions of a subject's hand (e.g., finger lengths). Note the photo is taken on a calibrated background.

jects used to ergonomic keyboards or alternate layouts (e.g., DVORAK) may have substantially different typing rhythms as compared to subjects used to using a standard keyboard with a QWERTY layout. Collecting this information permits us to account for these factors in our analyses.

Physical data

Physical data consists of measurements of a subject's height and weight, along with photographs of our subjects' hands and typing posture. A subject's electrocardiogram can be affected by his or her height and weight (or more precisely, body mass index, which depends on height and weight), so we take care to collect these data to allow us to properly read the electrocardiogram. A subject's height is taken via a measuring tape affixed to a wall, while weight is taken via a LifeSource digital scale (model number UC-322) (Manuals Online, 2018). Subjects are asked to remove their shoes and empty their pockets prior to the measurements to improve the accuracy of the measurements.

Photos of the subjects' hands and typing posture were also collected. These photos enabled us to ascertain a subject's hand geometry and body position while typing. Prior work in our lab has indicated that hand geometry and typing posture may have significant impact on a subject's typing rhythms. It seems conceivable that these may also impact manifestations of stress in typing; as a result, we take care to collect this information in the event that it may be useful in our analysis.

Hand photos are taken of both the left and right hand. Prior to each photo, the subject's hand is marked with blue dots, as depicted in Figure 5.2. Dots are placed on the knuckle of each finger, as well as the wrist of the subject. Photos are taken against a calibrated grid, enabling us to ascertain

26

the precise geometry of each subject's hands. Prior work in our lab has shown that hand geometry can significantly affect a subject's typing. For example, as compared to subjects with large hands, subjects with small hands tend to have much larger latencies between distant keys typed by the same hand.

Side photos of a subject's typing posture are collected by use of a still camera, positioned 90 cm away from the subject's chair. Prior to taking each photo, colored stickers are affixed to the subjects' shoulder, elbow, and hip. Care is also taken to ensure that the subject's ear hole is visible; subjects with long hair are asked to tie it back for the photo. Photos are taken while the subject is engaged in a typing task, ensuring that the posture captured reflects the subject's actual typing posture and not just a subject's estimation of his/her posture. These photos are taken in a manner that permits application of the Keyboard-Personal Computer Style instrument (K-PeCS) (Baker and Redfern, 2005).

Video data

Video data are collected using four Microsoft Life Studio Pro webcams positioned around the subject. Each of the cameras is capable of capturing 1080p resolution at 30 frames per second. These particular cameras were chosen because of their ability to record 1080p at 30 fps and also because they are equipped with threading in their base which permits easy attachment to professional photographic mounting hardware. The cameras are mounted using arms manufactured by Manfrotto. One end of the arm is clamped to shelving in the room for stability; the cameras are attached to the other end of the arm via the aforementioned threading. All four cameras are connected to the same machine, and video data are recorded using four instances of the Open Broadcasting Software (OBS) program (Open Broadcasting Software, 2018); one instance is used for each camera. An extra PCI-E USB card (StarTech.com, 2018) was used in the machine to permit all four cameras to simultaneously feed data to the same machine. In initial testing, we discovered that the data rate from four separate cameras was sufficient to overload the motherboard's default USB bus. To circumvent this problem, two webcams are attached to the motherboard's USB bus while the other two are attached to the PCI-E USB card.

Two of the cameras are positioned to the left and right of the keyboard, with camera hovering a few inches off the surface of the table. The purpose of these cameras is to capture a subject's hand positions from a side view during typing. A third camera is positioned directly above the keyboard. This camera is placed as low as possible while still being outside the peripheral vision of even a very tall subject. This camera's purpose is to capture the positioning of the subjects' fingers on the keys as they type. A fourth and final camera is directed at the face of the subject. This camera aids the experimenter in determining whether the subject is engaged with the present task and also provides a record of the facial expressions of the subjects. Such information can be helpful in ascertaining the affective state of the subject and may also explain anomalies in collected data (e.g., a sneeze may cause spikes in collected physiological data). Additionally, subjects are informed that they would be recorded throughout the experiment; this acts as a form of evaluation that is a potent source of perceived stress (Dickerson and Kemeny, 2004).

In addition to its purpose in the present experiment, the facial data was collected with the idea to use the collected data to assess the reliability of available facial recognition software. The collected data would be unusual in that it is accompanied by independent measures of the subjects' affective states at the time the data were collected. To make the collected data easier to process for such software, both the background and lighting were carefully controlled. A Lastolite Chromakey 1.8

x 2.75m green background (i.e., a greenscreen) was positioned behind the subject's chair to ensure a clear distinction between the subject's face and the background. The standard room lighting was turned off during the course of the experiment; illumination was provided by 2 Genaray SP-AD35 SpectroLED 9 lights. The LEDs were always set to the same setting – the third notch, approximately 60/100 strength – to ensure consistent lighting.

5.3.4   Subject relaxation

Throughout our experiment, we are interested in encouraging a relaxed, neutral state during two different rest periods. The materials used for achieving this state are handled differently for these two periods. We had originally planned to use the same technique for both sessions. However, feedback from subjects during pilot testing indicated that repeating the same technique twice actually caused feelings of discontent due to boredom. Some subjects even reported they felt stressed from being exposed to the same relaxation technique twice; clearly, this would be highly undesirable in our experiment.

In both periods, subjects watch video clips of underwater nature scenes of the Great Barrier Reef set to soothing music. Two clips are used – one for the baseline period and one for the recovery period. All subjects viewed the same clip in the baseline period and in the recovery period. Both clips are taken from the same source (Hannan, 1999).

During the first rest period, subjects are given pencil and paper and are asked to write down all animals they see in the video clip. During the second rest period, subjects are asked to pay attention to breathing and take breaths that are as long and deep as possible. These tasks are intended to prevent subjects from ruminating on topics unrelated to the experiment.

5.3.5   Stress induction

Stress induction is performed utilizing a combination of a multi-tasking framework and social evaluation – negative judgement from another person – from our experimenter. We initially planned on using software manufactured by Purple Research Solutions (Figure 5.3) for our framework, which effects a cognitive workload through a multitasking exercise presented as a game (Purple Research Solutions, 2014). The software has been demonstrated to induce stress in subjects (Wetherell and Carter, 2014). During the course of pilot testing this software, we discovered an issue with this software (detailed below). With permission, we decided to re-implement a version of this software that removed the identified issue; excepting this fix, our software is identical to the version manufactured by Purple Research Solutions.

Within the multi-tasking framework, the subject must engage 4 different tasks simultaneously. Points are awarded for good performance, and points are subtracted for poor performance. The subject's score is displayed in the middle of the screen.

In the upper-left quadrant is a set-membership task. The subject is shown a set of letters in the lower box (e.g., A, U, V, L, F, E, X, R), which disappear after a few moments. The subject is then shown a single letter in the circle (e.g., M) and must state whether the letter was a member of the set. Points are awarded for each correct answer; points are deducted for an incorrect answer or if the subject does not respond sufficiently quickly.

In the original software, we discovered an issue wherein subjects could rapidly accumulate points by clicking "False" in response to this task. This was caused by the fact that the letters in the circle were drawn (seemingly) uniformly at random from the 26 letters of the alphabet. As a consequence, letters were far more likely to be not included in the set. The scoring payout
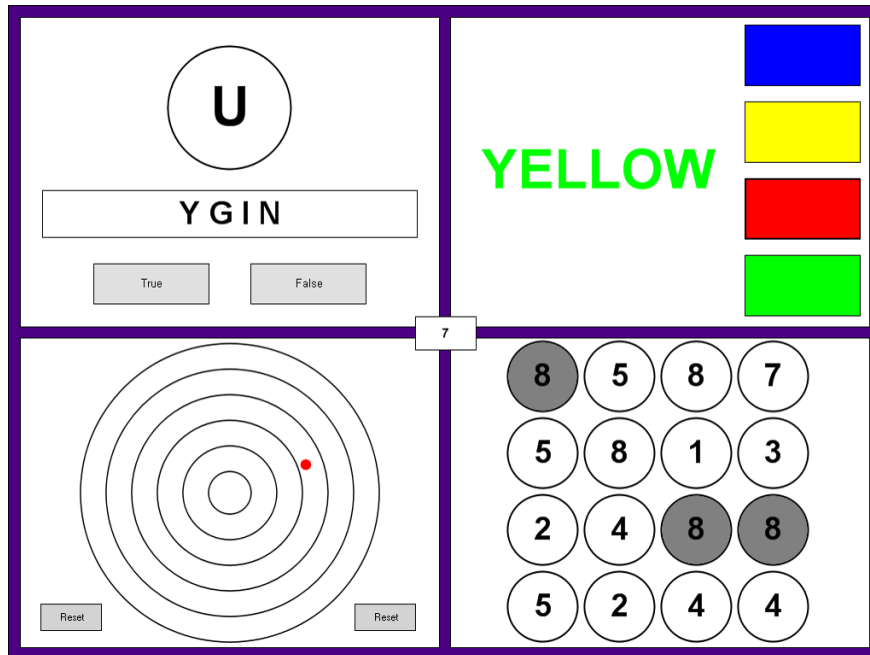
Figure 5.3: **Multi-tasking framework.** The subject must monitor four tasks concurrently (described in the body of the text). Points are awarded for good performance and subtracted for poor performance.

scheme – i.e., the values for a correct answer and incorrect answer – meant that subjects would be allowed to accrue points by always answering "False". During pilot testing, we discovered that this led to two major issues. First, subjects would often ignore the other three modules and focus on rapidly clicking "False" in this module as much as possible; this virtually eliminates the cognitive workload of the framework. Second, subjects would often have far too many points in the framework for our social evaluation to be effective; subjects with very high scores simply did not believe the administered social evaluation. In our version of the software, this issue was fixed by ensuring that the letter in the circle has a 50/50 chance of being contained in the set.

In the upper-right quadrant is a Stroop task (Stroop, 1935). The subject is shown the names of colors in different colored fonts (e.g., the word 'green' in a red font). The subject must click on the colored rectangle that corresponds to the color of the font, not the color spelled by the word. Points are awarded for each correct answer; points are deducted for an incorrect answer or if the subject does not respond sufficiently quickly.

In the bottom-left quadrant is a timing task. The red dot starts in the center of the concentric circles and drifts outward in a random direction. The subject can reset the position of the dot to the center of the circles by pressing the 'Reset' button. If this is done before the dot leaves the outer-most circle, points are awarded. The further out the dot is, the more points are awarded. If the dot leaves the outer-most circle, points are deducted.

In the bottom-right quadrant is a maximum-membership task. Each circle in the grid is randomly filled with one-digit numbers. Clicking on a circle highlights it in gray; clicking again removes the highlight. The subject must highlight all instances of the largest digit. Points are awarded when the subject successfully completes the task; point are deducted if a subject does not

29

complete the task sufficiently quickly.

5.4 Stimulus choice

Our choice of stimulus for this experiment was 280 repetitions of the phrase `great friends are good to have`, distributed across four sessions. 40 repetitions were collected in the warmup session, primarily to familiarize the subject with the experimental task and apparatus. The remaining 240 repetitions were collected in three sessions of 80 repetitions each; the three sessions were the baseline neutral, stress, and recovery.

In our study, as in many other keystroke dynamics studies, the question of stimulus selection arises. In most previous research, the selection process has been largely arbitrary. Researchers often pick out a phrase they deem convenient (e.g., the name of their institution) without giving considered thought to what phrase would be most suitable for the hypothesis entertained in the experiment; this results in different researchers using varied stimuli with no particular justification (Teh et al., 2013). In deciding the stimulus item for the present experiment, we wanted to avoid this arbitrariness. To do this, we developed a generic, principled method for selecting stimuli for affect-based keystroke dynamics research. We then applied this method to our present experiment, which resulted in the phrase `great friends are good to have`.

The full details of the stimulus selection process and the particular application to the present experiment can be found in Appendix A.3. Briefly, our goal was to find a phrase that was as easy to type as possible. Past experience in our lab suggested that easily-typed phrases tend to have lower practice effects and lower rate of errors than more difficult-to-type phrases. We devised a four-step process for stimulus selection and solicited opinions from 413 subjects from Amazon's Mechanical Turk to arrive at the chosen phrase. The four-step process was 1) to define the desired attributes of the end phrase, 2) to generate candidate phrases that fit these attributes, 3) to prune the pool of candidate phrases down, and 4) to experimentally determine which phrase best fit the list of desired attributes.

In the case of our experiment, the most notable requirements of the phrase was that it had to be memorable, devoid of emotionally charged text, and should be easy to type. In prior work in our lab, we had discovered that easily typed phrases tend to have lower variability in typing patterns, so we reasoned that we would maximize our ability to detect the effects of stress if we chose an easy-to-type phrase. We initially generated 100 phrases, which were then pruned down to 20 by removing the phrases that were least consistent with the requirements. These 20 phrases were then used as part of an experiment on Mechanical Turk. Subjects were repeatedly asked to type 2 of the 20 phrases and to indicate which one they thought was easier to type; phrases were chosen at random from the pool of 20. The resulting pairwise preferences were then transformed into an ordinal ranking using a Thurstone model. The highest-ranked phrase – the easiest one to type – was `great friends are good to have`, which was then taken to be the stimulus for the present experiment.

5.5 Power analysis

Our study is based on a single-subject design; adding more subjects does not help us to better identify physiological, psychological, and typing changes in a given subject. Rather, the primary advantage to increasing the number of subjects is to increase our ability to detect small clusters of subjects, with similar changes in typing rhythms, in the population. We estimate that 10 subjects

is the minimum required to meaningfully capture the inter-subject typing variance of subjects in the same cluster, which is critical for distinguishing a cluster from the remainder of the population. Thus, if a cluster comprises 1% of the total population, 1000 subjects must be run before we expect to see 10 members from the cluster; if it comprises 5% of the population, we need 200 subjects.

Due to the exploratory nature of the thesis work, it was decided that the smallest cluster we can reasonably explore is one that comprises 10% of the total population. A simulated binomial power analysis predicated on having at least an 80% chance of obtaining 10 subjects from such a cluster, determined that 124 subjects would be needed. To accommodate potential drop-outs and non-responders, our initial plan was to gather data from between 130 and 140 subjects.

5.6   Subject recruitment

A total of 132 subjects were recruited for the study through posters, posted on publicly accessible areas on campus, and through word-of-mouth. Our subjects are largely a convenience sample from a university campus – Carnegie Mellon University. Most subjects tend to be university undergraduates, with a mixture of graduate students, staff members, and persons outside of the campus community. Our original plans for this work included use of external research registries – e.g., the Center for Behavioral Decision Research – to supplement subjects recruited on campus and also to recruit a more diverse collection of subjects. However, we ultimately relied solely on recruitment from the campus community for two reasons.

First, we initially underestimated the number of subjects that we could recruit from campus through the use of posters and word-of-mouth recruiting. We had assumed that only about half of our recruitment needs could be met through such recruitment procedures, but ultimately were able to rely entirely on campus-based recruiting.

Second, we were concerned about the proportion of "seat-fillers" in our study. The term refers to subjects whose participation in the study is not grounded in good faith. Seat-fillers are typically individuals who seek to join studies for the sole purpose of making money and are more interested in providing responses that expedite the completion of the experiment rather than truthful responses. As an example of an issue, a seat-filler may not bother to ask clarifying questions about the experimental tasks despite not understanding what s/he needed to do in that task. This could lead to contaminated data; most genuine subjects could be expected to ask such clarifying questions.

A copy of the recruitment poster used can be found in Appendix A.4. Posters requested that the subjects contact our experimenter directly to confirm eligibility and to set up an appointment.

Subjects were asked to confirm their eligibility for the experiment by responding to the below list of questions. The expected responses from the subject is 'Yes' (or 'True') to all questions. A subject was required to meet all eligibility items to be eligible for the experiment.

1. I am at least 18 years old.
2. I speak English fluently.
3. I have at least three years of experience typing on a computer.
4. I can type at least 30 words per minute. (Typing at 30 words per minute means you can type the sample text below in 1 minute.)
5. I do not have any history of cardiac disorders.
6. I do not have any history of neurological disorders.
7. I do not have any history of anxiety or stress disorders.

8. I have never had a stroke.
9. I am not currently being treated by a doctor for a sleep disorder.
10. I do not suffer from any form of color-blindness.
11. My blood pressure is BELOW 140/90. (If you have a blood pressure ABOVE 140/90, you suffer from hypertension.)

Subjects were also informed that they would not be eligible for the study if they:

1. Consumed any alcoholic beverages within the 48 hours leading up to the experiment.
2. Consumed more than 3 caffeinated beverages within the 24 hours leading up to the experiment.
3. Consumed any caffeine or other stimulants within the 2 hours leading up to the experiment.
4. Consumed any psychoactive drugs, such as anti-depressants, Ritalin, marijuana, or LSD, within the 48 hours leading up to the experiment.
5. Had heard anything about the experiment outside of what was on the recruitment poster.

The purpose of asking this set of questions is to ensure that subjects are:

1. Legally eligible for the study
2. Able to properly understand and complete the study tasks in a reasonable amount of time
3. Do not have any medical conditions that may impact the validity of the study
4. Will exhibit his/her normal stress response in the study

In cases where subjects did not understand the question or did not know the answer, further clarification was provided by the experimenter. By far the most common issue was from subjects who did not know their blood pressure. Assuming all other questions were answered satisfactorily, and the subject stated that s/he was in general good health, an appointment was made for such subjects with the admonition that his/her blood pressure would be checked prior to the start of the experiment and that they would not be able to participate if his/her blood pressure was not in the required range; no subjects were rejected from the experiment due to such blood pressure concerns.

The experimenter was instructed to not schedule appointments for any individuals whose demeanor, responses, or statements seemed suspect in any way. Out of the 237 subjects that contacted us, 105 were rejected for some reason. Most commonly, subjects simply stopped responding to our e-mails. Other subjects could not find a suitable date or time for an experimental session or simply no-showed for their session. Finally, a small number of subjects did not meet the eligibility criteria.

All subjects were scheduled for the afternoon appointments. Since cortisol follows a diurnal cycle (Lovallo, 2005), we wanted to avoid time-of-day as a possible confounding variable. By running all subjects in the afternoon, they will all be in roughly the same portion of their diurnal cortisol cycle.

5.7   Experimental design

We have chosen to use an ABA single-subject design for the present experiment; this is also referred to as baseline-condition-baseline or baseline-condition-recovery. Simply put, the subject will provide a typing sample in a neutral baseline, in an induced stress condition, and then again in a neutral baseline (or recovered) state. The objective in using this design is to best enable us to

attribute any changes in typing behavior to stress and only stress, as we can use each subject as his/her own control.

We had originally considered using a simpler design, such as an AB or baseline-condition design. However, the limitation of such a design is that it is difficult to attribute any particular shift in typing, between the baseline and stress conditions, to stress and only stress. For example, changes in typing could be simply attributable to increasing fatigue or hunger, and not to stress. With an ABA design, if such external causes were responsible for typing changes, we would expect to see even greater shifts in the recovery session as the magnitude of the external cause would continue to increase over time; if stress were actually responsible for changes in typing, we would expect the recovery session to be quite similar to the initial baseline session.

We have also used a single-subject design for this experiment, wherein each subject's stressed typing data is primarily compared his or her own baseline typing data; that is, each subject acts as his/her own control. This enables us to capture individualized responses to stress that would not be possible if all data were aggregated. Of course, we can perform this aggregation at the analysis step where it is appropriate.

5.8 Experimental protocol

The protocol for our experiment ranges from the pre-experiment checks performed by our experimenter to the post-experiment cleanup tasks. In this section, we provide not only the timings of the events in the experiments, but also the specific instructions provided to the subjects. Moreover, where relevant, we discuss the reasons behind the choices made in designing the protocol and where alterations were made following pilot studies.

We start with a discussion of the protocol document and operations manual, generated to aid in the successful conduct of the experiment, and then proceed through the protocol in chronological order.

5.8.1 Protocol document and operations manual

In designing the present experiment, we noted that it would be several orders of magnitude more difficult and complex than experiments previously conducted by our experimenter. Prior experiments conducted in our lab were short (15-20 minutes) as compared to the duration of the present experiment (2-3 hours). While video and still photos were sometimes taken and occasionally a demographic survey would be administered to subjects, the present experiment contains a multitude of forms and additional equipment for capturing physiological measures, on top of the video and still photos. Prior experiments were also largely unconcerned with the affective state of the subject; we were happy to take subjects in whatever state they were in and merely made note of any extenuating circumstances that may have influenced typing. In the present experiment, we are highly concerned with the precise affective state of the subject; every time we made a decision regarding the protocol, we always evaluated how this decision might affect our ability to successfully control the affective state of the subject.

Given the relative complexity of this experiment, we felt that the probability of success would be significantly lowered if we asked our experimenter to execute the protocol from memory, or even with the assistance of a simple checklist. To address this, and inspired by checklists used by airline pilots, we opted to craft a protocol document that lists all of the steps required to successfully execute the experiment. The document totals 55 pages; Figure 5.4 shows a small snippet of this

29.0	**EKG electrode placement**	
29.1	Attach the EKG electrodes to the subject:	
29.1.1	*First, I am going to attach EKG electrodes on you like this picture.* *[Gestures to picture on wall.]*	

29.1.2	Steps:
	
	 1. *Please* **stand up**.
	
	 2. Put on a pair of **disposable gloves**.
	
	 3. Clean oils on skin with **alcohol wipes**
	
	     (a) **Boniest part of the right shoulder.**
	     (b) **Left lowest rib, 1 inch to the left of your nipple.**
	     (c) **Area below right rib, 1 inch to the right of your nipple and 2 inches below the left electrode position.**
	
	 4. Let the **alcohol dry a bit**.
	
	 5. Attach the electrodes[14].
	
	 6. *You can sit down now.*
		Complete? ☐

Figure 5.4: **Protocol snippet.** A small section of the experimental protocol checklist is depicted here. This particular section concerns the attachment of the ECG electrodes to the subject's body.

protocol. The document is broken down into 42 sections, each containing a single, logical task. It is organized in a 3-column fashion. The first column contains the step number, the middle column contains the specific action the experimenter must perform, and the final column contains checkboxes for the experimenter to mark that the action has been completed. Italicized text is used for things the experimenter must say, while bolded text is used to emphasize critical actions that must be performed.

Significant care was taken in crafting the protocol document, which went through dozens of iterations over a period of months. Significant feedback on the protocol document was provided by all researchers, including the experimenter herself. Our intention is that this protocol document will aid and assist any researchers who may wish to replicate or extend our work.

In addition to the protocol document, an operations manual was also written for the experiment. This 35-page document outlines the standard operating procedures for each of the software and hardware components involved in the experiment; more importantly, it also includes troubleshooting steps for any foreseen problems that may occur during the course of the experiment. The intent of this document is to allow our experimenter to remedy potential problems without requiring another researcher to be present and with only modest impact on an ongoing experimental session.

5.8.2 Pre-experiment setup

The experimenter's tasks begin approximately 30 minutes prior to the scheduled arrival of the subject. The chief goals in the setup stage are to ensure that all required materials are present, to

prepare a set of experimental materials (e.g., forms) for the subject, and to ensure that all required software and hardware work appropriately. Note that these checks typically take less than 10 minutes to perform; the additional time is intended for recovery if something is found to be amiss or if the subject arrives early.

Nearly all materials and equipment used in the experiment are dedicated to its exclusive use; thus it is highly unlikely that any items will be missing. Nevertheless, a check is made to ensure that the experiment can proceed without issue. Of particular concern in this check is to ensure that the three items requiring batteries – a remote control for the still camera, a remote control for the lights, and the scale for measuring the subject's weight – function properly; a check for extra batteries is equally paramount. To aid in the materials check, the protocol document lists them in order from right to left in the room; this permits the experimenter to simply sweep the room once instead of having to bounce back and forth. This was done to minimize the chance of omitting a check; we were particularly concerned with such an omission because it is so unlikely that any materials will be missing, possibly lulling our experimenter into a false sense of security.

The second setup stage is to assemble the experimental materials into a manila folder. This folder includes a copy of the protocol document – to be checked-off as the experiment proceeds – and sufficient copies of all the forms that will be administered during the course of the experiment (consent form, demographic survey, long-form STAI, PSS-10, short-form STAI, NASA-TLX). The experimenter also dates each form and records the subject number.

The final setup stage involves starting each piece of hardware and software used in the experiment to ensure it functions properly. This includes the webcams, PowerLab and LabChart, MTP, multi-tasking framework, the digital camera for the hand photos, the still camera for the K-PeCS photos, the scale, and the studio lights. Should any issues be noted, the experimenter uses the troubleshooting section of the operations manual to address any issues. To avoid any data-confusion issues, checks are performed using a fake subject number starting with a 't' (usually t000 or tdemo). This enables any data generated during these checks to be easily discarded later in the data-analysis phase.

5.8.3 Briefing and documentation

Following the arrival of the subject, the experimenter disconnects the phone in the room and places a "Do not Disturb" sign on the door. This is intended to minimize the possibility of distractions during the course of the experiment. The subject is also directed to turn his/her cell phone off, place the contents on their pockets in a cardboard box, and discard any chewing gum or mints in his/her mouth. The box is positioned to the right of the subject on the table; the box is within sight of the subject at all times and is impossible for the experimenter to reach without crossing close behind the subject (there is typically less than 6 inches between the back of the subject's chair and the greenscreen backdrop). Again, the goal here is to remove any potential distractions during the course of the experiment while ensuring that subjects do not feel uncomfortable that their possessions are out of sight.

Once settled, subjects are briefed about the purpose of the experiment and asked to provide their informed consent. As part of the consent process, the experimenter checks for the inclusion and exclusion criteria for the subject (see Section 5.6). Recall that the inclusion and exclusion criteria were previously communicated to the subject as part of the recruitment process; nevertheless, a second check is executed as part of the consent procedure since some of the criteria are short-term in nature (e.g., no caffeine consumption within the past 24 hours). The consent procedure also

involves a blood pressure reading to ensure that it falls within the eligibility range.

Once informed consent has been obtained, the subject is asked to fill out the demographic survey, long-form STAI, and the PSS-10. These are presented one at a time to the subject. Hand photos are also taken of both the subject's hands and the subject's height and weight are measured.

In total, the briefing and documentation procedure takes roughly 30 minutes. The actual time taken depends largely on how fast the subject reads through the consent form materials and on how many questions they have regarding interpretation of various items in the administered forms.

5.8.4 Familiarization period

The next stage of the experiment is the familiarization period. As its name suggests, the purpose of this period is to allow the subject to become familiar with the software and equipment used in the rest of the experiment as well as the visual analogue scales. This familiarization process is critical for two reasons.

First, we are interested in the effects of stress on a subject's typing behavior. These effects will be made more difficult to detect if a subject's typing is also being influenced by practice. During the experiment, subjects are using a keyboard that is almost certainly different from keyboards they use in their day-to-day life. Moreover, the MTP software is novel to the subjects. Without providing a familiarization period, we would risk conflating the effects of practice with the effects of stress. The issue of practice is discussed in detail in Section 7.2.

Second, we will be utilizing induced affective states during the course of the experiment; such states naturally decay over time. Any delays in the main body of the experiment will be detrimental to its success rate. We do not wish to have subjects asking lengthy questions about how to perform the experimental tasks in the middle of an induced affective state.

The familiarization period lasts roughly 20 minutes, beginning with an explanation of the visual analogue scales: the short-form STAI and NASA-TLX. It continues with a warmup typing task – subjects are asked to type the phrase "great friends are good to have" a total of 40 times. As in the proper experiment, repetitions of the phrase must be typed without error; the MTP software automatically greys out the text box and resets it whenever a typographical error is made. This typing task allows the subject to become familiar with the keyboard and the MTP software, while also allowing them an opportunity to become practiced at typing the phrase; the phrase is the same as the one typed during the main body of the experiment. During this warmup task, the experimenter takes still side photos of the subject's typing posture by remotely triggering a still camera.

The familiarization period concludes with an explanation of the four quadrants of the multi-tasking exercise. Subjects are encouraged to ask questions if they are at all confused, since their understanding of the task is vital later in the experiment. Once a subject feels s/he has understood the task, the experimenter starts a 2-minute warmup session of the multi-tasking exercise. The demands on the subject during this warmup session are significantly lower than during the main body of the experiment; the goal here is just to confirm the subject's knowledge of the task, not to induce stress. The subject is made aware that the task is intended to be extremely easy to ensure that they will focus on the task during the main body of the experiment. Once the subject concludes the 2-minute warmup session, they are given the opportunity to ask more questions or to participate in additional warmup sessions until they feel comfortable with the task.
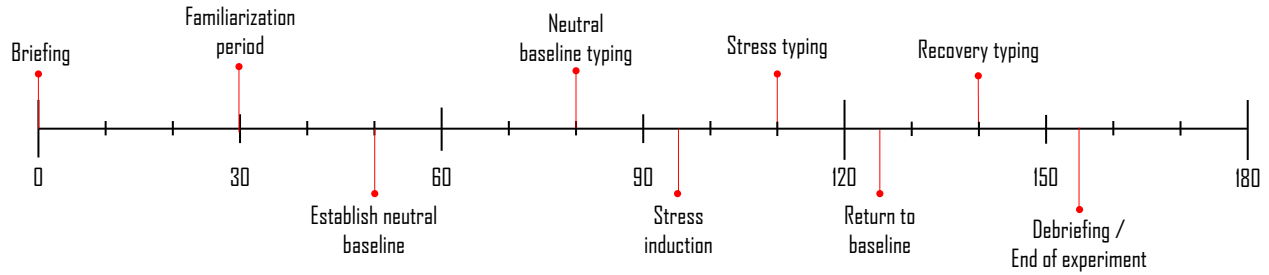
Figure 5.5: **Experimental Timeline.** A timeline, in minutes, of the major events in the experiment. Please note that the two rest periods for establishing and returning to baseline and the stress induction period are fixed in length, but all other events have variable length based on the pace of the subject.

5.8.5 Main experiment body

Upon completion of the familiarization period, the main body of the experiment begins. This is where the subject is attached to the array of physiological measurement sensors while they perform typing tasks in a neutral, then stressed, then neutral affective state. Before attaching the subject to the equipment, we strongly encourage him/her to use the restroom. Subjects are told that they will not be paid the full amount for the experiment if it is discontinued at any point before the end, including needing to use the restroom. The primary issue with a subject pausing the experiment in the middle to use the restroom is not the fact that equipment needs to be disattached and reattached; rather, the issue is that any induced affective state will be diminished or destroyed during the pause.

Figure 5.5 lays out the timeline for the main body of the experiment. Note that the duration of the rest periods and stressor exercise are exact; all other times are approximate and will depend on the subject's speed at completing tasks.

Sensor attachment. The first event in the main body of the experiment is the attachment of sensors. The ECG electrodes are attached first, in the positions described in Figure 5.1, while the subject is standing. Once these electrodes are attached, the subject is asked to sit down. The respiration belt is then attached around the waist of the subject.

The ECG wires are then attached to the electrodes; the experimenter asks the subject to snap the leads onto the electrodes. It is not only easier for the subject to do this than for the experimenter to do this, but this also lessens any discomfort the subject might feel about having his/her personal space invaded. To ensure proper lead attachment, the experimenter listens for the snapping sound associated with a lead attaching to the electrodes; the experimenter also gently tugs on the leads to ensure firm attachment. Following lead attachment, the blood pressure cuff is then attached to the subject. All associated wires are then taped to the desk using heavy masking tape and/or duct tape. Taping these wires is important for several reasons. First, the wires can occasionally get in a subject's way if they are not taped. In some extreme circumstances, it could be possible for a subject to roll over the wires with the wheels of the chair; this is obviously highly undesirable. Second, the wires are actually relatively heavy; in early pilot experiments, we found that the weight of the wires could contribute to ECG leads detaching from the electrodes or the electrodes themselves peeling off the subject's skin. Third, movement in the wires is transmitted back into the electrodes. Thus, an accidental strike of the wires could lead to a disrupted ECG or respiration signal for a few seconds until the wires stop moving.

Following attachment of all the equipment, the experimenter checks to see that signals are coming through properly on LabChart. This involves setting the captured range of the ECG and respiration signals – the experimenter aims for a range that is roughly twice as wide as the normative signal coming from a resting subject. The subject is asked to take several deep breaths during this time to help the experimenter in determining an appropriate range for the respiration signal. Note that there is an inherent tradeoff in setting the range for the signals. Setting a wider range lessens the possibility that the signal will go out-of-bounds at some point in the experiment – say, due to a subject violently coughing or sneezing. However, this lowers the resolution of the signal. Setting a narrow range increases the resolution of the signal, but increases the chance of an out-of-bounds event. The range differs from subject to subject, with the physiology of the subject (primarily the amount of tissue and muscle in the upper chest) being the primary factor for HRV and the snugness of the respiration belt being the primary factor for respiration. Differences in range can largely be attributed to a subject's physiology (e.g., a larger subject may have more mass between the heart and the ECG electrodes, diminishing the captured electrical reading) and, in the case of respiration, the precise tightness of the attached equipment.

Establish neutral baseline. Once the experimenter is satisfied that the physiological equipment is properly attached, she initiates the start of the first rest period. This period is 30 minute in length, as measured by a stopwatch. During these 30 minutes, the subject is asked to perform a simple task while watching a relaxing video of an underwater scene (Hannan, 1999). The subject is handed paper and pen and asked to write down all unique animals that s/he sees while watching the video; this simple task is intended to keep the subject's attention on the experiment, so that s/he does not ruminate on non-experiment-related thoughts. The experimenter also asks the subject to try to relax as much as possible while performing this task.

In our initial pilot studies, we had asked our subjects to merely sit in the chair for the duration of the 30 minutes in silence. Early pilot subjects reported that this was sufficiently boring that they had difficulty relaxing. We then augmented this rest period by playing music in the background – we chose Pachelbel's Canon (Pachelbel, 1680) as it is generally regarded in the literature as being a relaxing piece of music. The music was looped in the background for the 30 minute duration of the rest period. Despite the view in the literature that this is a relaxing piece, we found that subjects were generally annoyed by the repetitiveness associated with looping the music; subjects who were trained musicians were particularly vehement in their dislike of listening to 30 minutes of Pachelbel's Canon. Needless to say, this did not achieve the desired relaxation effect. To avoid the repetitiveness, we switched to playing the aforementioned video. Unfortunately, we found that subjects often fell asleep during the video; since we do not want to capture any "grogginess" effect in our subjects, this was not desirable. By adding a simple distractor task – writing down all unique animals found in the video – we found that subjects did not fall asleep and also self-reported that the 30 minutes were relaxing.

During the course of the first rest period, blood pressure readings are taken every 5 minutes. The first reading is taken right at the beginning of the session, so a total of 7 readings are taken. ECG and respiration data are taken continuously throughout the course of the rest period.

Upon the conclusion of the first rest period, the first set of VAS forms (short-form STAI and NASA-TLX) are administered. These self-reports are designed to serve as a point of comparison for the remainder of the experiment. It is expected that subjects will self-report high levels of relaxation and very low workload experienced during the 30-minute rest period.

Baseline typing sample. Immediately after the VAS forms have been administered, the subject

is asked to provide the first neutral typing sample. This consists of 80 repetitions of the phrase "great friends are good to have". No blood pressure readings are taken during the course of the typing sample; in pilot studies, we found that the natural movements involved in typing routinely caused the blood-pressure meter to struggle to obtain a reading. This would often cause the cuff to significantly inflate, often to the point of significant physical discomfort for the subjects. In some cases, the Critikon would fail to obtain a reading altogether, triggering a noisy alarm. Due to these issues, we opted to not take blood pressure readings during the course of the typing sample. Both ECG and respiration data are still continuously collected, however.

After the first typing sample has concluded, the second set of VAS forms is administered.

Stress exercise. Once the first neutral typing sample has concluded, the stress exercise begins. The subject is instructed to move the keyboard out of the way and place the mouse in a comfortable position. Once this is done, the experimenter begins a 15-minute multi-tasking exercise task; this is far more strenuous than the warmup task the subject has previously experienced. Throughout the exercise, the experimenter applies social evaluation to the subject; care is taken to apply a steady stream of social evaluation to the subject regardless of how well s/he is doing. This involves pointing out mistakes that the subject has made – whether this is in the form of incorrect responses or allowing a module to timeout – and pressuring the subject to work faster. Our experimenter also reported that repeatedly sighing and standing immediately behind the subject were also effective at increasing a subject's stress level. Some form of negative evaluation – either a prompt to work more quickly, a statement that the subject is not performing well enough, or a bout of sighing – was administered once per minute. Additionally, at the 5-minute and 10-minute mark in the exercise, the experimenter harangues the subject for not completing the task quickly or accurately enough. The subject is also warned that they need to significantly improve their performance if they wish to receive the full compensation for the experiment.

As with the rest period, blood pressure readings are taken every 5 minutes throughout the exercise, with the first reading taken right as the exercise starts. In contrast to the typing sample, we are able to take blood pressure readings during the multi-tasking exercise because the cuff is attached to the subject's non-mouse arm. Movement in the non-mouse arm is fairly minimal during the exercise, whereas typing involves use of both hands. In addition to blood pressure data, ECG and respiration data are collected throughout.

Once the multi-tasking exercise has finished, the third set of VAS forms is administered. Extreme care is taken at this stage to shorten the time period between the completion of the multi-tasking exercise and the beginning of the next typing session, as the induced stress will fall off over time.

Stressed typing task. Before beginning the stressed typing session, the subject is asked to move the mouse out of the way and return the keyboard to a comfortable typing position. The stressed typing task consists of 80 repetitions of the phrase "great friends are good to have" – same as in the neutral typing sessions.

Once the typing sample is complete, the fourth set of VAS forms is administered.

Second rest period. Prior to starting the second rest period, the subject is informed that they performed well enough in the multi-tasking exercise to receive full payment, provided that they complete the remainder of the experiment to the experimenter's satisfaction. This is designed to help reduce the levels of stress in the subject in preparation for the second rest period.

The second rest period consists of a 15-minute rest, while the subject watches additional scenes from the underwater movie. In lieu of a written exercise, we asked our subjects to focus on taking

long, deep breaths throughout the course of the 15 minutes. In pilot studies, we found that offering subjects a different task – i.e., not asking them to write down animals in the video again – reduced boredom and helped in lowering stress. A shorter, 15-minute, rest period also seemed more effective than a 30-minute rest period; we suspect that the longer rest period gave subjects an increased sense of boredom and frustration, inhibiting a return to a relaxed state.

As with the first rest period, blood pressure readings are taken every 5 minutes. ECG and respiration data are collected continuously throughout this time.

Once the rest period is complete, the fifth set of VAS forms is administered.

Recovery typing sample. Once the rest period is over, the third and final typing sample is collected from the subject. As before, the sample is 80 repetitions of the phrase "great friends are good to have". Once the typing sample is complete, the VAS forms are administered for a sixth and final time.

Sensor detachment and debriefing. When the final VAS forms are completed, the experimenter informs the subject that the experiment is now complete. The sensors are then removed from the subject and the experimenter answers any questions the subject may have concerning the experiment.

The experimenter also inquires whether the combination of the multi-tasking framework exercise and social evaluation was effective at stressing out the subject. Any comments that the subject may have about the experiment are also taken into consideration, in case improvements could be made to improve future runs of the experiment.

Following this debriefing and payment, the subject departs.

5.8.6 Clean-up

A few tasks remain for the experimenter before the conclusion of the experiment from her point-of-view. Most importantly, all of the collected data must be archived. Then, all equipment is turned off and cleaned with alcohol wipes, where applicable. The blood pressure records are then printed out and all materials from the experiment are placed into the manila folder.

A check is then performed to ensure that sufficient consumable items (e.g, electrodes, tissues) are available for upcoming experiments.

Should there be any pressing issues regarding the experiment (e.g., subject was rejected for some reason or equipment malfunction), the experimenter notifies the researchers immediately so that adjustments can be made prior to the arrival of the next day's subject.

5.9 Instructions to subjects

We break down the instructions to the subjects by task.

Familiarization period – general. Subjects are informed that the purpose of this period is to allow them to become comfortable with the experimental equipment (i.e., keyboard, mouse, chair) and the important experimental tasks. Subjects are instructed to adjust the chair and keyboard so that they are in a comfortable position. This includes adjusting the chair height, armrests, and keyboard position. Subjects are asked to place their hands in a typing position to ensure that the equipment is in a comfortable position. All subjects were also given the option of using a footrest; this was mainly applicable for shorter subjects whose feet may not reach the floor given the desk height.

Multi-tasking framework. During the familiarization period, subjects are instructed to become familiar with the multi-tasking framework used in the experiment. Written instructions are provided to the subject (see Appendix A.5). These instructions are a bulleted, written description of the tasks described in Section 5.3.5. The experimenter also verbally describes the tasks as the subject reads through the written document. The reason for providing both a written and verbal instructions for the task are to maximize the chances that a subject understands the way the framework works. After completing a 2-minute warmup task for the multi-tasking framework, subjects are encouraged to ask any questions they may still have. They are informed that if they have any doubts or uncertainties about the tasks, they should raise doubts at this point, as the experimenter will not be able to answer questions once the main experiment begins.

During both the familiarization period and prior to the stress exercise, it is emphasized to the subjects that the framework is points driven – their goal is to accrue as many points as possible – and that they will be awarded points for correct answers while having points deducted for incorrect answers or missed responses. It is particularly emphasized to the subjects that they must be as fast and accurate on ALL of the tasks in order to achieve as high a score as possible; we placed extra emphasis on this after we observed subjects tending to focus on only one or two modules in our pilot studies.

First rest period. Prior to the first rest period, subjects are instructed to make themselves as comfortable as possible in the chair. They are instructed to relax as much as possible, while completing the simple task of writing down all animals in the video. Subjects were instructed to simply identify the broad category of animal, as though they were describing the animals to a small child (e.g., fish, shark). Since the intent of this task is simply to stop subjects from ruminating on non-experiment-related thoughts, we wanted to avoid subjects worrying about identifying the specific species of animal.

Typing samples. Prior to each typing sample, subjects were asked to ensure that the chair and keyboard were still comfortable for them. For all typing samples, subjects were instructed to type at a normal pace. It was stressed to subjects that this was neither an accuracy contest nor a race, and that it is their natural typing rhythm that we are interested in.

VAS forms. For both VAS forms (short-form STAI and NASA-TLX), subjects were instructed to make a single vertical mark when responding to the questions. We discovered in pilot testing that some subjects had a tendency to "color-in" their response by making multiple vertical marks. This made it quite difficult to score the forms, so we explicitly asked our subjects to make a single vertical mark.

Second rest period. Prior to the second rest period, subjects were instructed to once again make themselves as comfortable as possible in the chair. Again, they were instructed to relax as much as possible. Subjects were asked to control their breathing by making their breaths as long and deep as possible.

Chapter 6

Question 0: Did the stressor work?

Before we dive into the three major questions of the thesis, we must first answer a critical question: did the stressor work? The objective of this thesis is to ascertain the extent to which an affective state – stress – can be detected through changes in typing rhythms. It would not be meaningful to address this objective without first considering whether our subjects were in the expected affective states when they provided their typing samples.

If our stressor was successful, we would expect to see significant changes in the physiological and psychological measures we collected in the experiment. Specifically, we would expect to see elevated levels of blood pressure, decreased heart-rate variability measures, and elevated psychological inventory scores. We would also expect to see subjects exhibiting return-to-baseline behavior in the recovery session. That is, with the stressor removed, we expect subjects to revert to their initial baseline scores on the collected measures.

We start with an analysis of the physiological and psychological measures that most directly answer the question of whether the stressor worked. We will explore these changes both in aggregate and also examine the most extreme cases. We then present a brief aggregate analysis of the typing data, leaving the detailed analysis to the next three chapters – one for each major question in the thesis.

6.1   Aggregate changes in physiological and psychological measures

Of the data that we have collected, the data that most directly answer the questions about the affective state of our subjects are the physiological and psychological measures.

The physiological data analyzed are the heart-rate-variability (HRV) measures – consisting of the median R-R interval and SDRR (standard deviation of R-R interval) – and the blood pressure measures – consisting of systolic and diastolic blood pressure, mean arterial pressure, and pulse rate. As previously mentioned in Chapter 4, we excluded respiration data from analysis because of significant issues from movement artifacts and because we issued directions to our subjects to explicitly control their breathing during the second rest period.

The psychological measures include responses to the short-form STAI and the NASA-TLX, which measure anxiety and workload, respectively.

HRV measures are computed for 5-minute intervals in each of the baseline, stress, and recovery periods. For each measure, we then average the values for all intervals in a period to arrive at an

average value for that period. Blood pressure measures for each period are likewise obtained by averaging all readings within that period.

For the STAI and NASA-TLX data, a subject's responses to the inventories after the first rest period, the stressor exercise, and the second rest period were used as the baseline, stress, and recovery period values, respectively.

6.1.1 Statistical testing

The objective of our statistical testing is to determine whether there are statistically significant changes in the physiological and psychological measures between the baseline, stress, and recovery sessions. A particular concern we wish to mitigate is the effect of performing multiple statistical tests. Doing so requires upward correction of all obtained p-values, based on the number of tests performed, to avoid inadvertently increasing the false-detection rate; most simply this can be done with a Bonferroni correction (Miller, 1981).

The easiest way to mitigate the effects of multiple testing is to reduce the number of statistical tests run. Our objective is to run the broadest possible tests first. If these tests indicate that there is no significant result, we can stop at that point; by stopping early, we do not run further statistical tests, which lowers the required Bonferroni correction. If the broadest tests are significant, then we proceed with the next round of tests, which are more focused. Again, if this round of tests are not significant, we can stop early. If they are significant, we proceed to the most detailed round of tests. One can picture this process as an inverted-pyramid scheme. The broadest test is run first and each significant test causes us to run a narrower test until a test is either not significant or the narrowest possible test has been run.

In the context of our analysis, the broadest possible test is a MANOVA (multivariate analysis of variance). When a MANOVA is applied to a given set of measures, we are examining whether there are any changes in any of the measures between the neutral, baseline, and recovery conditions. If a MANOVA is not significant, we are assured that none of the measures in the set significantly shifted between the neutral, baseline, and recovery conditions; a non-significant MANOVA rules out any changes in any of the measures. On the other hand, a significant MANOVA does not reveal the extent of the changes. It could be that all measures significantly vary between all the conditions or it could be only a single measure varying between two of the three conditions.

Once a significant MANOVA has been obtained, we run a series of ANOVAs (analysis of variance) to hone in on which changes occurred. An ANOVA examines whether a particular measure (e.g., systolic blood pressure) differs significantly between the baseline, stress, and recovery conditions. If an ANOVA is not significant, we are assured that the measure did not vary significantly between any of the three conditions. A significant ANOVA, however, does not indicate which of the conditions differed. It could be that all three sessions differ from each other significantly or it could be only two out of the three conditions that differ significantly.

When a significant ANOVA is obtained, we proceed to the most detailed series of tests, a set of four paired t-tests for a given measure. The t-tests compare 1) baseline vs. stress, 2) stress vs. recovery, 3) baseline vs. recovery, and 4) combined baseline and recovery vs. stress to determine which of these conditions are significantly different from one another. The paired t-tests are the most fine-grained analysis of this data that we can statistically perform. They indicate whether or not the measure differed significantly between two specific conditions.

6.1.2 MANOVAs

We performed three different repeated-measure MANOVAs, one for 1) HRV data, 2) blood pressure data, and 3) NASA-TLX data; no MANOVA is performed for the STAI data as the instrument produces a single score. We opted to conduct three separate MANOVAs instead of a single, large, joint MANOVA because the large number of covariates required for fitting the joint MANOVA would have been problematic both statistically and computationally. Moreover, conducting three separate MANOVAs allows us to identify which particular types of data differ between neutral and stressed states.

For each set of measures, Mauchly's test for sphericity was performed (Maxwell and Delaney, 2004, p.542). Sphericity is an underlying assumption for a repeated-measures MANOVA. It refers to the equal variance assumption when examining the differences between each of the time points at which measures are taken. In our case, there are three time points: 1) initial neutral baseline, 2) stressed, and 3) recovery neutral baseline. When this assumption is violated, the MANOVA results can be inaccurate; this can be easily remedied through either a Greenhouse-Geisser correction (Maxwell and Delaney, 2004, p.543), which adjusts the degrees of freedom in the MANOVA to compensate for the violation of sphericity. Our goal in checking for sphericity is merely to determine whether we need to undertake the correction. In almost all the analyses, we found that sphericity was violated – Mauchly's test was highly significant ($p < 0.001$). In the instances where it was violated, we employed the Greenhouse-Geisser correction. As an analogy, one can think of this process as checking whether a required pre-condition is met (e.g., Does the car have gas?); if the pre-condition is not met, then a correction must be taken (e.g., fill up the tank).

There are a number of statistics used to assess the significance of MANOVAs. The four most common statistics are Pillai's statistic, Wilks' statistic, Lawley-Hotelling's statistic, or Roy's greatest root (Maxwell and Delaney, 2004, p.721). For the sake of brevity and clarity, we restrict ourselves to presenting our results with Pillai's statistic. The nature of our results does not change if any of the other three statistics are used.

6.1.3 ANOVAs

A repeated-measures ANOVA was performed for each physiological and self-reported measure. There are 13 such measures: four blood pressure readings, 2 HRV readings, 1 STAI score, and 6 NASA-TLX scores. The purpose of the ANOVAs is to examine whether there are significant differences between the baseline, stress, and recovery sessions. Note that a significant ANOVA test does not reveal which of the three sessions are different, merely that at least two of them are different. To pinpoint the sessions that are different, we turn to paired t-tests.

6.1.4 Paired t-tests

Four paired t-tests were performed for each physiological and self-reported measure. The purpose of the t-tests is to allow us to determine precisely which sessions are significantly different from each other. The four tests compare: 1) the initial baseline vs. stress sessions (AB), 2) the recovery baseline vs. stress sessions (BA), 3) the two baseline sessions against each other (AA), and 4) the average of the baseline and recovery sessions vs. the stressed session (ABA).

As we are performing numerous paired t-tests (13 measures $\times$ 4 tests/measure = 52 tests) in each analysis, we need to employ a correction factor to ensure that the desired false discovery rate is respected. In this work, we use a significance value of $\alpha = 0.05$, which directly corresponds to a

false discovery rate of 5%. The simplest way to ensure the false discovery rate remains at 5% is to divide the obtained p-values by the number of statistical tests conducted. This technique is known as a Bonferroni correction (Maxwell and Delaney, 2004, p.202). Because we perform 52 tests per analysis, we obtain a corrected significance level of $\alpha' = 0.05/52 = 0.000962$. As such, we only claim statistical significance when $p < \alpha' = 0.00962$, instead of when $p < \alpha = 0.05$, as would be typical.

6.1.5 Results

Figure 6.1 shows the mean physiological and self-reported measures for subjects in the baseline, stress, and recovery sessions. Note that all measures are expected to increase with stress with the exception of HRV measures and Performance (NASA-TLX), which are expected to be lower. It is further expected that recovery values will not constitute a full return to baseline.

The results almost entirely meet expectations. All measures increase/decrease as expected between the initial baseline and the stressed sessions. The measures then decrease/increase toward the initial baseline values, but do not make a full return. The sole exception is in SDRR, where subjects actually experience a higher SDRR in the recovery session than in the baseline session. A possible explanation for this phenomenon is that subjects are asked to perform a deep-breathing exercise for the entirety of the second rest period; in the first rest period, subjects are asked to watch a video and perform a trivial task (write down all the animals that appear).

MANOVAs for blood pressure, HRV, and NASA-TLX data were all highly significant ($p < 0.001$), indicating that there are significant changes in at least one measure in each group between at least two of the three conditions (baseline, stress, recovery).

For all three MANOVAs, Mauchly's test for sphericity was significant (blood pressure: $p < 0.001, \epsilon = 0.85$, HRV: $p = 0.002, \epsilon = 0.92$, NASA-TLX: $p = 0.001, \epsilon = 0.90$). Accordingly, we use a Greenhouse-Geisser correction when conducting the MANOVA. Results were significant for blood pressure (Pillai's trace= $0.84, F(1.70, 96.9) = 390.86, p < 0.001$), HRV (Pillai's trace= $0.67, F(1.84, 104.88) = 143.16, p < 0.001$), and NASA-TLX (Pillai's trace $= 0.94, F(1.80, 101.7) = 1026.50, p < 0.001$).

Table 6.2 shows the p-values resulting from the paired t-tests; the significance results are summarized in Table 6.3. Note that every single measure is highly significantly different under AB, BA, and ABA comparisons. This strongly indicates that our experiment was successful, in aggregate, as our subjects' states are different when comparing neutral and stressed conditions. Also note that some measures are significantly different between the two baselines while others are not. For the measures that are not different, it would appear that subjects properly returned to baseline after the administration of the stressor. In aggregate, it appears that subjects did not fully return to baseline after the stressor. However, since all tests were highly significant for the BA comparison, we can conclude that there was some return to baseline, just not a complete one.

6.2 Identifying potential non-responders

So far, we have seen that the aggregate changes in physiological and psychological data are entirely within our expectations. While this supports our claim that our subjects, in aggregate, were in the appropriate affective states when they provided their typing samples, it does not directly address the claim that every single subject was in the appropriate affective states. It could be plausible, for example, that a large percentage (say, 90%) of subjects did respond, but that the remainder were
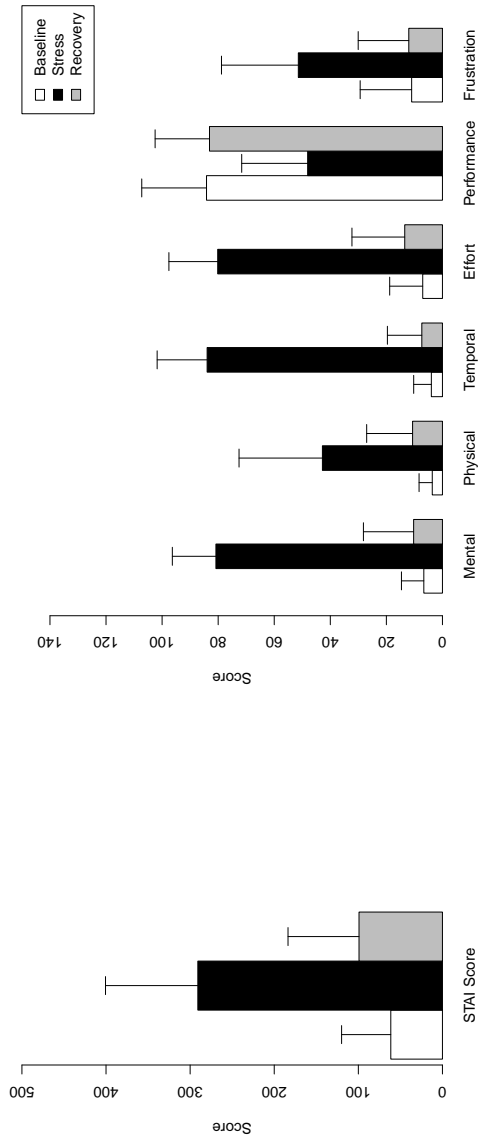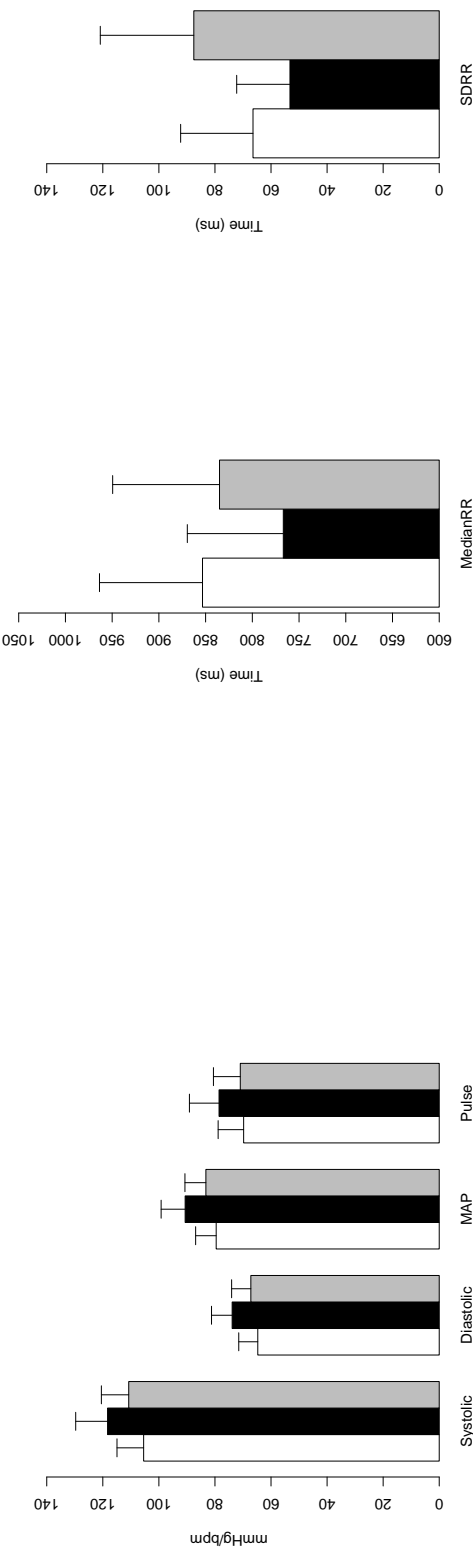
Figure 6.1: **Baseline, stress, and recovery comparison.** Mean physiological and self-reported measures for subjects in the baseline, stress, and recovery sessions. Error bars represent one standard error. Please note that the y-axis for the median R-R plot is truncated. It is expected that all measures will increase with stress with the exception of HRV measures and Performance (NASA-TLX), which are expected to be lower with stress.

Measure	DF	F-statistic	p-value < 0.001?
Sys	2/230	300.98	✓
Dia	2/230	251.93	✓
MAP	2/230	358.65	✓
PR	2/230	107.30	✓
Median R-R	2/230	101.97	✓
SDRR	2/230	152.85	✓
STAI	2/230	370.81	✓
Mental	2/230	1205.45	✓
Physical	2/230	162.25	✓
Temporal	2/228	1523.55	✓
Effort	2/230	837.49	✓
Performance	2/230	117.80	✓
Frustration	2/230	162.83	✓

Table 6.1: **ANOVA results.** Degrees of freedom, F-statistic, and p-values for repeated-measure ANOVAs run on each physiological and self-reported measure. Note that one subject (s206) was omitted from the Temporal ANOVA due to missing data. Note that all obtained p-values were below 0.001, indicating highly significant changes for each measure.

not responsive to the stressor. Such non-responders could, for example, be seat-fillers who are merely interested in collecting the compensation for the study without actually making an honest attempt to participate.

In a perfect world, we would address this issue directly, in one of two ways. First, we could have some well-defined set of thresholds that define whether a subject is neutral or stressed. For each subject, we could then ascertain whether that subject was in the appropriate states at the relevant points in the experiment. Unfortunately, subjects have widely varying individual differences in their physiological and psychological manifestations of stress. Consequently, to the best of our knowledge, such thresholds do not exist.

Second, we could attempt to perform a statistical analysis to demonstrate that each subject has statistically significant physiological and psychological differences between neutral and stressed conditions. Unfortunately, this too is not possible. The issue here is that we have very few data points on each measure for each subject. For example, the STAI and NASA-TLX forms are administered once in each condition, providing a grand total of one data point per subject. Performing a meaningful statistical analysis would require at least a small handful of data points for each subject. This would, in turn, require us to have brought subjects back for multiple sessions, vastly expanding the experiment beyond our available resources.

With the two direct approaches ruled out, we must instead use an indirect approach. This approach relies on establishing two facts. First, that even low-responding subjects still responded to stress. Second, that there are no subjects who are consistent low-responders across all measures. If these two facts hold, we are forced to conclude that all subjects responded on some measures to the stressor.

To establish that low-responding subjects still responded to stress, we repeat the statistical analyses that we have just showcased, but focus on the lowest responding quartile for each measure.

Measure	AB	BA	AA	ABA
Sys	7.17×10^{-44}	1.17×10^{-24}	1.32×10^{-23}	1.62×10^{-37}
Dia	1.27×10^{-42}	1.14×10^{-27}	1.95×10^{-9}	5.17×10^{-39}
MAP	3.91×10^{-48}	4.45×10^{-31}	1.24×10^{-18}	3.81×10^{-43}
Pulse	4.17×10^{-24}	9.88×10^{-19}	2.92×10^{-2}	1.14×10^{-23}
Median R-R	2.69×10^{-25}	6.71×10^{-16}	8.09×10^{-4}	1.71×10^{-22}
SDRR	5.34×10^{-12}	3.06×10^{-28}	9.47×10^{-21}	2.42×10^{-24}
STAI	3.53×10^{-43}	3.47×10^{-38}	2.98×10^{-8}	7.96×10^{-43}
Mental	2.18×10^{-78}	2.74×10^{-61}	1.05×10^{-2}	2.96×10^{-72}
Physical	4.46×10^{-28}	2.52×10^{-22}	2.10×10^{-6}	2.30×10^{-26}
Temporal	1.27×10^{-74}	4.18×10^{-69}	2.43×10^{-3}	1.45×10^{-73}
Effort	1.25×10^{-68}	3.78×10^{-56}	6.39×10^{-4}	4.40×10^{-67}
Performance	9.24×10^{-23}	1.03×10^{-22}	6.28×10^{-1}	4.99×10^{-25}
Frustration	9.67×10^{-27}	4.28×10^{-26}	5.63×10^{-1}	5.03×10^{-28}

Table 6.2: **Paired t-test p-values.** P-values resulting from paired t-tests for each physiological and self-reported measure. Comparisons are made between the initial baseline and stressed sessions (AB), between the recovery baseline and stressed sessions (BA), between the two baseline sessions (AA), and between the average of the two baseline session and the stressed session (ABA). See Table 6.3 for a list of outcomes. While it is more common to report p-values as being below 0.001, in lieu of reporting the actual number, we report the actual obtained p-value to underscore the highly significant nature of our tests.

For example, we identify the lowest 25% of responders, as measured by systolic blood pressure, and repeat our battery of MANOVAs, ANOVAs, and paired t-tests to establish this group still responded to the stressor. This will be repeated for each of the 13 measures we have collected.

To establish that there are no consistent low-responders across all measures, we perform a rank-based analysis of our subjects' responses. On each measure, we will rank our subjects from most-responsive (rank 1) to least-responsive (rank 116) – that is, largest change to smallest change. For each subject, we will then compute the average rank across all measures. We will show that no subject has an average rank in the lowest 25%; that is, we do not have any subjects who are consistent low-responders across all measures. Rather, even if a subject shows little or no response to one measure, s/he will have a significant response in other measures.

6.2.1 Analyzing the lowest responders

We must start by defining what is means for a subject to be low-responding on one of the 13 measures of interest. In lieu of defining a strict threshold for response, we take a more pragmatic approach by defining the 25% of the subject population that showed the lowest response on a given measure to be the low-responding subjects for that measure. As there are a total of 116 subjects, there are exactly $116 * 0.25 = 29$ low-responding subjects in the bottom 25%.

For each of the 13 measures of interest, we then repeat the statistical analyses presented in Section 6.1 using only the 29 lowest-responding subjects for that measure. We start examining the results within a given measure. That is, we ask the questions: Do the 29 lowest-responding subjects, as identified by systolic blood pressure, still show a statistically significant change on

Measure	AB	BA	AA	ABA
Sys	HS	HS	HS	HS
Dia	HS	HS	S	HS
MAP	HS	HS	HS	HS
Pulse	HS	HS	NS*	HS
Median R-R	HS	HS	S	HS
SDRR	HS	HS	HS	HS
STAI	HS	HS	S	HS
Mental	HS	HS	NS*	HS
Physical	HS	HS	S	HS
Temporal	HS	HS	NS*	HS
Effort	HS	HS	S	HS
Performance	HS	HS	NS	HS
Frustration	HS	HS	NS	HS

Table 6.3: **Test result summary.** Results of paired t-tests for each of the physiological and self-reported measures. Comparisons are made between the initial baseline and stress sessions (AB), between the recovery baseline and stress sessions (BA), between the two baseline sessions (AA), and between the average of the two baseline sessions and the stress session (ABA). A significance value of $\alpha = 0.05$ was used; after a Bonferroni correction, the corrected significance level is $\alpha' = 0.05/52 = 0.000962$. Entries in the table are either NS (not significant at the 0.05 level), NS* (not significant, but only after the Bonferroni correction), S (significant after Bonferroni correction), and HS (highly significant, $p < 10^{-10}$). The pilot subjects are dropped in this analysis, leaving us with 116 subjects; all data points in a session are used for BP and HRV data.

systolic blood pressure? How about when median R-R is used instead of systolic blood pressure? How about STAI scores? And so on, for all 13 measures.

Table 6.4 shows the results of the analysis for each of the 13 measures. The left-most column lists the measure used to identify the 29 lowest-responding subjects. The MANOVA column contains the significance results for the MANOVA applied on the group containing that measure. For example, for systolic blood pressure, the MANOVA is performed using the 4 blood pressure measures since systolic blood pressure is a blood pressure measure. Likewise, for median R-R, the MANOVA is performed using the 2 HRV measures. The ANOVA column contains the significance results for the ANOVA applied on the measure in the left-most column. If systolic blood pressure is used to identify the lowest-responding subjects, the ANOVA is performed on systolic blood pressure. Finally, the four right-most columns depict the t-test results, on the measure, between the listed conditions. If systolic blood pressure is used to identify the lowest-responding subjects, then the paired t-tests are performed on systolic blood pressure.

Note that, without exception, all of the MANOVAs are still significant even when examining only the 25% lowest-responding subjects. 11 of the 13 ANOVAs are still significant; the exceptions are pulse rate (PR) and Frustration. Likewise, excepting pulse rate and frustration, all of the neutral vs. baseline t-tests are still significant. Therefore, we can conclude that for at least 11 of the 13 measures, even the lowest-responding subjects still responded.

We now turn our attention to pulse rate and frustration, which had non-significant ANOVAs

Measure	MANOVA	ANOVA	AB	BA	AA	ABA
Sys	< 0.001	< 0.001	< 0.001	0.013	< 0.001	< 0.001
Dia	< 0.001	< 0.001	< 0.001	< 0.001	0.543	< 0.001
MAP	< 0.001	< 0.001	< 0.001	< 0.001	0.002	< 0.001
PR	< 0.001	0.036	0.670	0.054	0.056	0.099
Median R-R	< 0.001	< 0.001	< 0.001	< 0.001	0.004	< 0.001
SDRR	< 0.001	< 0.001	< 0.001	< 0.001	0.002	< 0.001
STAI	N/A	< 0.001	< 0.001	< 0.001	0.032	< 0.001
Mental	< 0.001	< 0.001	< 0.001	< 0.001	0.224	< 0.001
Physical	< 0.001	< 0.001	< 0.001	0.001	0.467	< 0.001
Temporal	< 0.001	< 0.001	< 0.001	< 0.001	0.545	< 0.001
Effort	< 0.001	< 0.001	< 0.001	< 0.001	0.921	< 0.001
Performance	< 0.001	< 0.001	< 0.001	< 0.001	0.104	< 0.001
Frustration	< 0.001	0.281	0.647	0.161	0.286	0.238

Table 6.4: **P-value results for the lowest-responding quartile.** For each measure, the 25% lowest-responding subjects were identified. Then, we repeated the MANOVA, ANOVA, and paired t-test evaluations using only this smaller subset. For example, in the first line, the 25% lowest-responding subjects were identified, as measured by systolic blood pressure. Using this subset, we then repeated the statistical analyses. A MANOVA was conducted for the blood pressure measures (as systolic blood pressure is a blood pressure measure), an ANOVA was conducted for systolic blood pressure, and finally 4 paired t-tests were conducted for systolic blood pressure. Note that aside from PR (pulse rate) and Frustration, all baseline vs. stress comparisons are significant. This means that even the lowest-responding 25% of subjects still responded. PR and Frustration are examined further, as detailed in the text.

and neutral vs. stress t-tests. What we have observed is that, in aggregate, the 25% of lowest-responding subjects on pulse rate do not significantly respond on pulse rate and the 25% of lowest-responding subjects on frustration do not significantly respond on frustration. One might wonder: Do these two sets of subjects respond on other measures?

The answer is that they do. Table 6.5 presents the t-test analyses for the lowest 25% of respondents, as measured by pulse rate. For simplicity's sake, we present only the t-test analyses. With the exception of median R-R, which is very closely tied to pulse rate, the other neutral vs. stress t-tests are significant. This indicates that this group of subjects responded to the stressor on almost all measures collected in the study.

Likewise, Table 6.6 presents the fuller set of analyses for the lowest 25% of respondents, as measured by frustration. Note that all of the performed neutral vs. stress t-tests are statistically significant.

6.2.2 Rank-based analysis

We have seen in the previous section that isolating the 25% of lowest responders, by each measure, still results in significant changes on all other measures. We can therefore conclude that any subject who is outside of the 25% of lowest responders (i.e., in the top 75% of responders), must have responded to the stressor.

50

Measure	AB	BA	AA	ABA
Sys	< 0.001	0.001	< 0.001	< 0.001
Dia	< 0.001	< 0.001	0.004	< 0.001
MAP	< 0.001	< 0.001	< 0.001	< 0.001
PR	0.6697	0.0543	0.0562	0.0987
Median R-R	0.0238	0.1569	0.4919	0.0307
SDRR	< 0.001	< 0.001	< 0.001	< 0.001
Score	< 0.001	< 0.001	0.0037	< 0.001
Mental	< 0.001	< 0.001	0.1741	< 0.001
Physical	< 0.001	< 0.001	0.0442	< 0.001
Temporal	< 0.001	< 0.001	0.3525	< 0.001
Effort	< 0.001	< 0.001	0.1474	< 0.001
Performance	< 0.001	< 0.001	0.8622	< 0.001
Frustration	< 0.001	< 0.001	0.3305	< 0.001

Table 6.5: **Paired t-test p-values, lowest 25% of responders as measured by pulse rate.** The lowest 25% of responders, as measured by pulse rate, were identified. Then, 4 paired t-tests were conducted for all 13 collected measures. Note that all neutral vs. stress tests are significant, excepting those for pulse rate and median R-R; these two measures are closely tied as they both measure heart beat rate. These results indicate that the lowest 25% of responders, as measured by pulse rate, still demonstrate a significant stress response by 11 of the 13 measures.

The question remains: On average, are all subjects actually outside of the 25% of lowest responders? If this were the case, then we can conclude that all subjects must have responded to the stressor. To assess this, we start by rank-ordering our 116 subjects from most responsive (rank 1) to least response (rank 116) on each of the 13 measures (systolic blood pressure, median R-R, STAI score, etc.). We then compute the average rank for each subject over the 13 measures. Note that a (hypothetical) subject who is the most responsive on every measure would have an average rank of 1, while a subject who is the least responsive on every measure would have an average rank of 116. The cutoff for the bottom 25% of responders is rank 87, since that subject must be in the top 75% (116 * 0.75 = 87).

Figure 6.2 depicts the average rank for our subjects. Note that all subjects have an average rank above 87. That is, there are no subjects who are consistently in the bottom 25% of responders across all measures. The lowest rank belongs to subject s287, who has an average rank of 80.5.

6.3 Aggregate changes in typing measures

So far in this chapter, we have explored the changes in our subjects' psychological and physiological measures when exposed to stress. We have conducted analyses to conclude that our subjects did respond to the stressor, in the expected manners. Having seen this, we eagerly turn our attention to the changes in typing measures.

While we will defer the deep-dive analyses to the next three chapters of this thesis, we now perform a cursory examination of the changes in our subjects' typing. For the purposes of this section, we restrict our attention to three measures: average hold time, average keydown-keydown latency, and number of errors made for each correctly-typed repetition (recall that subjects are
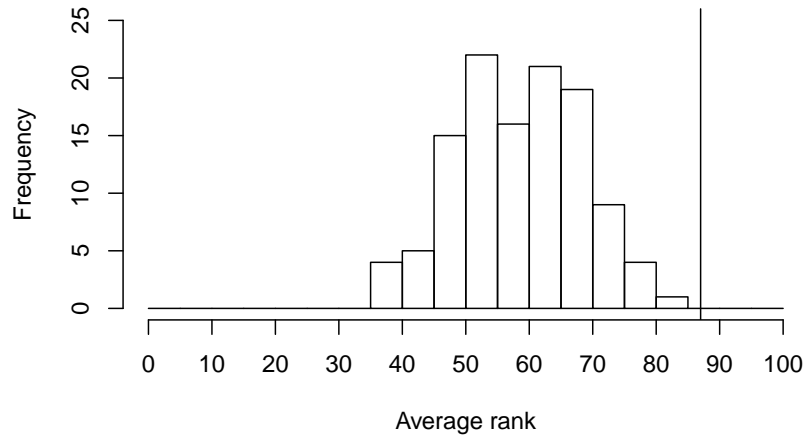
Figure 6.2: **Average rank across all measures.** Subjects were ranked from 1 (most responsive) to 116 (least responsive) on each of the 13 physiological and psychological measures (systolic blood pressure, median R-R, STAI score, etc.). The average rank over all 13 measures was then taken for each subject. Note that all subjects have a ranking less than 87 (vertical line), which is the cutoff for the lowest 25% of responders (116 * 0.75 = 87). This indicates that there are no consistently low-responding subjects.

Figure 6.3: **Typing measure changes.** Mean hold times, keydown-keydown (DD) latencies, and errors per correctly typed repetition across all subjects in the baseline, stress, and recovery sessions. Error bars represent one standard error.
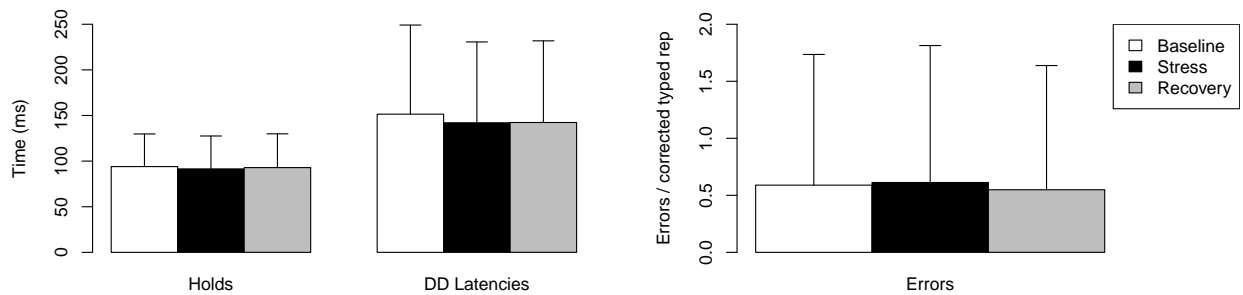
Measure	AB	BA	AA	ABA
Sys	< 0.001	< 0.001	< 0.001	< 0.001
Dia	< 0.001	< 0.001	0.0370	< 0.001
MAP	< 0.001	< 0.001	< 0.001	< 0.001
PR	< 0.001	< 0.001	0.8112	< 0.001
Median R-R	< 0.001	< 0.001	0.7260	< 0.001
SDRR	0.0023	< 0.001	< 0.001	< 0.001
Score	< 0.001	< 0.001	0.0462	< 0.001
Mental	< 0.001	< 0.001	0.2553	< 0.001
Physical	< 0.001	< 0.001	0.0119	< 0.001
Temporal	< 0.001	< 0.001	0.1775	< 0.001
Effort	< 0.001	< 0.001	0.0646	< 0.001
Performance	0.0146	0.0035	0.8642	0.0049
Frustration	0.6471	0.1606	0.2861	0.2377

Table 6.6: **Paired t-test p-values, lowest 25% of responders as measured by Frustration.** The lowest 25% of responders, as measured by Frustration, were identified. Then, 4 paired t-tests were conducted for all 13 collected measures. Note that all neutral vs. stress tests are significant, excepting those for Frustration. These results indicate that the lowest 25% of responders, as measured by pulse rate, still demonstrate a significant stress response by 12 of the 13 collected measures.

prompted to restart a repetition if they make a typographical error).

Our preconceptions about typing suggest that subjects would have markedly shorter hold and latency times, while also having a markedly higher number of errors when under stress. Figure 6.3 shows the changes in these three measures. Careful examination of the figure demonstrates that our preconceptions are correct, but it is also apparent that the aggregate effect size is small. In fact, the aggregate effect size is so small that statistical analyses are likely not relevant; even if these aggregate effects were statistically significant, they are not meaningful.

Having seen fairly sizable changes in our subjects psychological and physiological measures, it may come as a surprise that the typing measures seem to have hardly changed. We shall see that the aggregation – averaging over all subjects – has completely smoothed over the actual changes from stress. As we shall see in the next few chapters, there are indeed sizable changes in our subjects' typing as they are exposed to stress. However, these changes are highly individualized; the effect of our aggregation is to make these changes seem almost nil.

6.4   Summary

We started the experiment expecting to see particular changes in the collected physiological and psychological measures. It was expected that all blood pressure measures would rise with stress, all heart-rate variability measures would fall, the STAI score would rise, and the NASA-TLX measures would rise, excepting Performance which would fall. Using a combination of MANOVAs, ANOVAs, and paired t-tests, we found that these expectations were met. The expected changes were observed in the collected measures when subjects transitioned from a neutral to stress state. We also observed a partial recovery on every measure but SDRR; we attribute the unexpected SDRR results to the fact that we instructed our subjects to concentrate on taking deep breaths

during the second rest period.

In addition to observing the expected results in aggregate, we also saw that even the lowest-responding subjects still responded to the stressor. We established this by demonstrating that even the lowest-responding 25% of subjects, as given by any measure, still responded to at least 11 of the 13 measures. Moreover, we saw that there are not consistently low-responding subjects; no subjects are consistently in the lowest-responding 25%. Consequently, we are left to conclude that all subjects must have significantly responded to the stressor in some fashion.

Having established that the stressor was effective for our subjects, we turn our attention to a closer analysis of the collected typing data in the next chapter.

Chapter 7

Question 1: Identifying markers for stress on an individual level

We saw in the previous chapter that the experiment appeared to be successful – all subjects had the expected physiological and psychological changes between their neutral, stress, and recovery typing samples. We also saw, a bit surprisingly, that the changes in the typing data appeared to be fairly minor. In this chapter, we perform a closer examination of the typing data; we shall see that the diminished response was largely an artifact of aggregating over all subjects.

The three primary objectives for this chapter are to: 1) ascertain whether neutral and stressed typing from the same subject can be differentiated, 2) to confirm that it is actually stress that effects this differentiation and not a confounding variable like practice, and 3) to identify the markers that facilitate this differentiation. Recall that markers are features of a subject's typing data that differ significantly between neutral and stressed typing.

We start by attempting to differentiate neutral and stressed typing by employing machine learning algorithms. More precisely, we show that a small sampling of off-the-shelf machine learning techniques are able to reliably and successfully distinguish between neutral and stressed typing from any given subject. We then turn our attention to the issue of practice in the collected typing data; as previously alluded to in Chapter 5, this concern had a major influence on our experimental design. Finally, we will use more traditional statistical analyses to reveal the markers that facilitate successful classification.

7.1   Classification

We use two types of keystroke features in this thesis work: hold times (duration between the press and release of a given key) and keydown-keydown latency times (duration between the pressing of two consecutive keys). As our phrase contains 31 characters in total (including the Return key), we thus have 31 hold times and 30 latency times. If the presence of stress causes changes in typing, we would expect it to manifest in one or more of the keystroke features that we have collected.

As we saw at the end of Chapter 6, there are only minute changes in the aggregate hold and latency times, despite marked changes in aggregate physiological and psychological measures. It is natural to ask: what about changes on a individual level (i.e., not aggregated across all subjects)? We begin our exploration of this question by attempting to use machine learning (ML) algorithms to differentiate between neutral and stressed typing within the same subject. We opt to employ
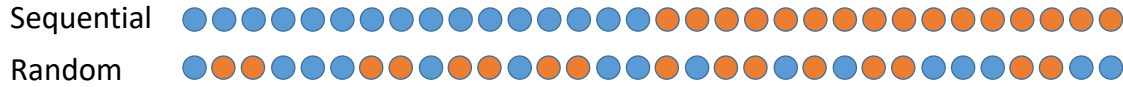
Figure 7.1: **Data selection methods.** The two employed methods for data selection in our evaluation. Blue dots represent training data and red dots represent testing data.

Classification task	Selection method	Baseline training/testing reps	Stress training/testing reps	Recovery training/testing reps
AB	Sequential	Reps 1-40/41-80	Reps 1-40/41-80	None/None
BA	Sequential	None/None	Reps 1-40/41-80	Reps 1-40/41-80
ABA	Sequential	Reps 1-20/21-40	Reps 1-40/41-80	Reps 1-20/21-40
AA	Sequential	Reps 1-40/41-80	None/None	Reps 1-40/41-80
AB	Random	40/40 random reps	40/40 random reps	None/None
BA	Random	None/None	40/40 random reps	40/40 random reps
ABA	Random	20/20 random reps	40/40 random reps	20/20 random reps
AA	Random	40/40 random reps	None/None	40/40 random reps

Table 7.1: **Classification regimes.** Repetitions used in the training and test set in each of the eight classification regimes used in this chapter. Note that the training set and test set are always disjoint and that all random draws are without replacement.

ML algorithms first because they are readily able to handle non-linear and multi-feature changes, even if we are not necessarily aware of the nature of these changes; statistical approaches, which we will also employ later, often require significant work to handle these complexities. Effectively, ML algorithms act as a more stringent filter. If there is something to be found, ML algorithms are more likely to reveal it than statistical approaches; if they do not find anything, it would provide strong evidence that there is simply nothing to find (i.e., there are no changes between neutral and stressed typing).

At a high level, the ML algorithms we employ all work in a similar fashion. The primary inputs to each algorithm are repetitions of neutral typing and repetitions of stressed typing from the same subject. These input data are referred to as the *training data*. Each algorithm then *learns* a *model* for neutral and stressed typing from this subject. Naturally, different algorithms will form different models, which may lead one algorithm to perform better than another on a given subject. To evaluate the goodness of these models, we provide each learned model with some *testing data*, consisting of new repetitions of neutral and stressed typing from the given subject. It is critical to note that the training and testing data are disjoint; the goodness of the learned models is thus contingent on their ability to *predict* whether never-before-seen repetitions from the subject were produced in a neutral or stressed state. This disjointness is critical. It ensures that the algorithm has actually learned something general about the subject's typing in these two states; it is not just regurgitating memorized answers.

7.1.1 Classification regimes

In this chapter, we will focus on eight different classification regimes. These are the cross-product of four classification tasks and two methods for selecting training and testing data. Classification

tasks are defined by the sessions of typing data that they compare. The four tasks are baseline neutral vs. stress (AB), stress vs. recovery neutral (BA), joint baseline and recovery neutral vs. stress (ABA), and baseline neutral vs. recovery neutral (AA). For each subject, we hope to see high *classification accuracy* (the rate of successful prediction) when comparing neutral and stressed sessions. We would hope to see lower accuracies when comparing the two neutral sessions against each other.

The two methods for selecting training and testing data are illustrated in Figure 7.1: sequential – where the training data and testing data are consecutive in the session – and random – where training data and testing data are drawn randomly from the repetitions in a session. We would expect that choosing consecutive training and testing data would result in lower accuracies than when choosing these data randomly. This is because any local trends in the data (e.g., a few consecutive repetitions with some unusual typing pattern) will be naturally accommodated by the random selection method, as there will likely be repetitions with this unusual behavior in both the training and testing set. Such trends will not necessarily be accommodated by the consecutive selection method, as all repetitions with the pattern may occur in only the training data or in only the testing data.

When comparing two sessions against each other (i.e., for AB, BA, and AA classification), we always use 40 repetitions of each session for training and 40 repetitions of each session for testing, regardless of the data selection method. With a sequential data selection method, we use the first 40 repetitions of each session for training and the remaining 40 repetitions for testing. With a random data selection method, 40 repetitions of each session are randomly chosen, without replacement, for training and the remaining 40 repetitions are used for testing. To alleviate any bias due to a particularly good or bad random draw, we average our results over 100 random draws. That is, 100 random draws will be conducted, 100 different models will be learned, and the presented results will be the average accuracy over the 100 resultant accuracies.

When performing joint baseline and recovery neutral vs. stress comparisons (ABA task), we use 20 repetitions of each of the neutral sessions and 40 repetitions of the stress session for training. This split ensures that the classifier is provided with an equal number of neutral and stress repetitions. Similarly, 20 repetitions of each neutral session and 40 repetitions of the stress session are used for testing. With a sequential data selection method, we use the first 20 repetitions of each neutral session for training, along with the first 40 repetitions of the stress session. The testing data are comprised of repetitions 21-40 in each neutral session and repetitions 41-80 for the stress session. With a random data selection method, we form the training data by taking 20 repetitions chosen at random, without replacement, from each neutral session along with 40 randomly-chosen repetitions from the stress session. The testing data are likewise formed, taking care to ensure that the training and testing data are disjoint. As with the other classification tasks, we present average results over 100 draws.

Table 7.1 contains a summary of the used repetitions for training and test data in each of the eight classification regimes.

7.1.2 Classifiers

We employ three classification algorithms (*classifiers*) in this chapter: 1) random forest (RF), 2) support-vector machine (SVM), and 3) lasso-regularized logistic regression (LASSO). These algorithms were chosen due to their readily-available off-the-shelf implementations, their relative simplicity, and their reputation for excellent performance on real-world problems.

Note that our goal in this work is not to find or create an algorithm that maximizes the classification accuracy on one or more of the eight classification regimes we have identified. Rather, we simply seek to demonstrate that sufficient differences exist between neutral and stressed typing to permit successful classification. We presume that any results obtained in this section could be improved upon with sufficient effort. In accordance with this philosophy, we have chosen to use the default settings – as chosen by the software packages we employ – for each of the three classifiers that we use.

We provide here a brief description of the three classifiers employed in this chapter. Our goal here is to merely convey the high-level idea behind each classifier. We would direct the interested reader to the cited materials for the detailed mathematics behind these classifiers.

Random forest. The random forest algorithm was first introduced by (Breiman, 2001). A random forest is comprised of many random decision trees. In order to train a single random tree, a bootstrap sample of size n is first drawn from the existing data with replacement, where n is the number of data points in the training data. The random tree is a decision tree which is trained by considering $\sqrt{p}$ randomly chosen features as candidates for each split, instead of all features (where p denotes the total number of features). The random forest classifier is then created by combining many random trees via a majority vote; we use 500 random trees in our work. We use the `randomForest` function within the `R` package `randomForest` (version 4.6-2) (Liaw and Wiener, 2002) in our implementation.

Support-vector machine. The current formulation of a support vector machine (SVM) was first introduced by (Boser et al., 1992). In an SVM, the objective is to find a hyperplane which not only separates the data, but which does so with a maximal margin. The margin is defined as the minimum distance between any data point and this separating hyperplane. Of course, it is not always possible to separate the data completely (as that would require data that could be classified perfectly), so typically a linear penalty, $C \times \psi$, is assigned, where C is a chosen constant and ψ is the amount by which a data point falls short of achieving the margin. A larger value of C more harshly punishes violations of the margin. The goal then becomes maximizing the margin while ensuring that not too many points violate the margin by too large an amount. For our version of the classifier, we employ the commonly-used RBF (radial basis function) kernel which permits non-linear relationships to be established and set $C = 1$, as is default setting for the `ksvm` function in the `R` package `kernlab` (version 0.9-25) (Karatzoglou et al., 2004).

Logistic regression. Logistic regression is a classical machine learning algorithm which still sees regular use. It presumes that the likelihood of a given data point being generated under stress is related to a linear, weighted "score" (akin to linear regression). It then applies a logistic function to turn these numeric scores to probabilities that a given data point was generated under stress. We use a slightly modified version with lasso (L1) regularization, which encourages small weights to be zero (i.e., removing certain keystroke features from the score) (Hastie et al., 2009, p.125). This is done by setting all weights with magnitude below some λ to be 0 while subtracting λ from all larger weights. In our implementation, we use the `cv.glmnet` function from the `R` package `glmnet` (version 2.0-13) (Simon et al., 2011); this function automatically selects the optimal value of λ prior to running the algorithm.

7.1.3 Evaluation procedure

Our evaluation procedure is fairly straightforward: we evaluate each classifier, under each classification regime, for each subject. For classification regimes that involve sequential data selection,

Classifier	AB	ABA	BA	AA
RF	**76.64**	**72.14**	**71.83**	**84.87**
SVM	73.67	69.66	70.04	81.06
Lasso	71.83	67.76	68.75	80.02

Table 7.2: **Sequential classification accuracies.** Classification accuracies are shown for each of the four classification tasks for each of the three employed classifiers. The data selection method was set to sequential to achieve these figures. Accuracies are aggregated over all subjects.

Classifier	AB	ABA	BA	AA
RF	**89.47**	**80.96**	**86.04**	**93.67**
SVM	85.56	76.79	82.06	89.68
Lasso	83.18	73.78	79.85	88.00

Table 7.3: **Random classification accuracies.** Classification accuracies are shown for each of the four classification tasks for each of the three employed classifiers. The data selection method was set to random to achieve these figures. Accuracies are aggregated over all subjects and over 100 random draws of data.

this means that we are learning 3 (classifiers) x 116 (subjects) = 348 classifiers for that regime. Accordingly, a total of 348 classification accuracies will be obtained. For a regime that involves random data selection, a total of 3 (classifiers) x 116 (subjects) x 100 (random draws) = 34,800 classifiers are learned for that regime. However, since we average over the 100 random draws, we still obtain 348 classification accuracies in total.

7.1.4   Results and discussion

We first examine the results for each classifier, under each classification regime, while aggregating over subjects. Table 7.2 depicts the accuracies for the sequential data selection method and Table 7.3 depicts the accuracies for the random data selection method. Note that a statistically significant result (i.e., beating chance) requires an accuracy of 60%. This value is obtained by forming a one-sided (greater) binomial 95% confidence interval with 80 total trials (the size of the test set) and a null hypothesis of 50% accuracy. The smallest number of successes (i.e., correct classifications) required for this confidence interval to not include the null hypothesis is the required number of correct classifications to obtain above-chance accuracy. In our particular case, 48 correct classifications are required, corresponding to 60% accuracy. Regardless of the classification regime or the classifier, we see that the average classification performance handily beats chance.

It is also worth noting that the Random Forest classifier uniformly dominates the other two classifiers. That is, regardless of the classification task or data selection method, it out-performs the other two classifiers. Accordingly, we will restrict our analysis for the remainder of the chapter to only the Random Forest classifier; results are similar, but slightly worse for the other classifiers.

In comparing the results for the two data selection methods, we see that the random method leads to markedly higher results than the sequential method. This is as expected, since the random method is much more able to deal with local changes in typing; that is, if a subject has a few consecutive repetitions with a common deviation from typical typing, this can be easily accommodated by the random selection method but not by the sequential selection method.
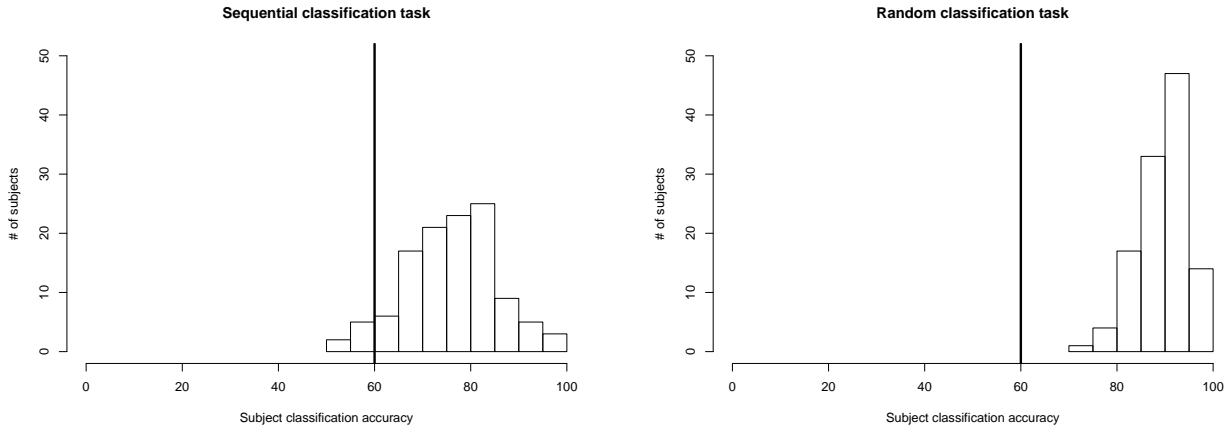
Figure 7.2: **Individual subject accuracies.** Histograms of accuracies for individual subjects, as obtained by the Random Forest classifier, are depicted for both the sequential (left) and random data selection method (right) for the AB classification task. Accuracies for the random data selection method are averaged over 100 random draws of training and testing data. Vertical lines at the 60% mark in both histograms denote above-chance classification rates. Note that 61 of our 116 subjects (more than half) have classification accuracies above 90% in the random classification task.

Another interesting observation is the manner in which the classification accuracies vary across tasks. As expected, accuracies for the ABA task are lower than either the AB or BA task. Additionally, the AB accuracies are also higher than the BA accuracies. This suggests that the recovery to neutral/baseline is incomplete – subjects are still slightly stressed despite undergoing a second rest period. This agrees with the general trend of physiological and psychological measures, as seen in Chapter 6. Quite surprisingly, we also see that the AA accuracies are considerably higher than all other classifications. That means that it is easier to discriminate between baseline and recovery typing than to discriminate between either neutral session of typing and stressed typing. This suggests there may be some sort of external factor, such as practice, that is responsible for this trend. For the moment, we will set this issue aside, but we will return to it shortly in Section 7.2.

Having seen that the overall accuracies, aggregated over all subjects, are quite good for Random Forest, we delve down into subject-level accuracies. We wish to know whether the classifier performed above chance on all subjects or if it merely excelled on some subjects while performing at or below chance levels for others. Figure 7.2 depicts the per-subject accuracies for the classifier for both sequential and random data-selection methods. We have chosen to depict the results for the AB classification task only, as we have already seen that subjects do not fully recover by the time the recovery typing sample is provided. Note that most classification accuracies are above the chance mark (60%) for the sequential data selection method and all are above this mark for the random data selection method. This confirms that above-chance classification is a general trend across all subjects, not merely an artifact of a few high-performing outliers. Also noteworthy is the fact that 61 out of 116 subjects (more than half) achieved a classification accuracy above 90% in the random classification task.

7.2 Ruling out practice as a potential confound

We saw in the prior section that there were unexpectedly high baseline vs. recovery (AA) classification accuracies. In fact, these accuracies are higher than those obtained when comparing data from one or both of the neutral typing sessions against the data from the stress typing session. This observation raises the question of whether a non-stress-related force is present in the data. The most prominent such force, and one that we heavily considered as a potential confounding variable (as noted in Chapter 5) is that of practice.

Practice, in the context of typing, is familiar to anyone who has had the "pleasure" of being assigned (or forced to choose) a new password. Initially, the password is difficult to type; it feels awkward, slow, and clumsy. With time, typing the password becomes increasingly easier, quicker, and more fluent, until it is comfortable to type. Through repetition – i.e., with practice – the user's typing changes significantly.

If practice effects were prominent in the collected typing data, it might fully explain the high AA classification accuracies. The classifier could effectively have learned to discriminate between typing at various levels of practice. Therefore, we now turn our attention to ruling out practice as a dominant effect in the collected typing data. We shall see that while some practice effects do manifest in the collected data, these effects are fairly minimal. Moreover, these practice effects can be accommodated – effectively "subtracted out" of the data – and doing so only has a minor impact on our ability to discriminate between neutral and stressed typing data. That is, even with the practice effect removed from the data, there remains a large signal in the data, which we attribute to changes in stress state.

7.2.1 Quantifying practice effects

To address practice in the collected typing data, we must first quantify the magnitude of practice in the data we have collected. Fortunately, the phenomena of practice is well studied in the social and behavioral sciences. A standard approach in this literature is to use practice curves (Ritter and Schooler, 2001) to numerically describe practice.

A practice curve is an exponentially decaying function, which can be described by the following equation:

$$y = M + B \cdot (x + E)^{-\beta}$$

which contains two variables, x and y, and four parameters (M, B, E, and β). Within the context of our typing data, this equation expresses the length of a keystroke feature, y, as a function of the number of repetitions of the phrase that a subject has typed, x. The four parameters also have direct interpretations.

M is the predicted minimum amount of time required for the subject to type the password (on average), even with an infinite number of practice repetitions. B is the "range of learning" which predicts the total difference (in seconds) in the length of a keystroke feature between a state of zero practice (i.e., fully unpracticed) and a state of infinite practice (i.e., fully practiced). E represents the prior experience of a subject before he began our typing task, cast in terms of the number of repetitions he has effectively already typed. A large value for E would indicate that a typist was already fairly practiced at the task, perhaps due to being a skilled touch-typist, while a small value for E indicates little prior experience. Finally, β is the learning parameter, which governs how quickly a subject learns. A large value of β indicates that the subject learns rapidly.
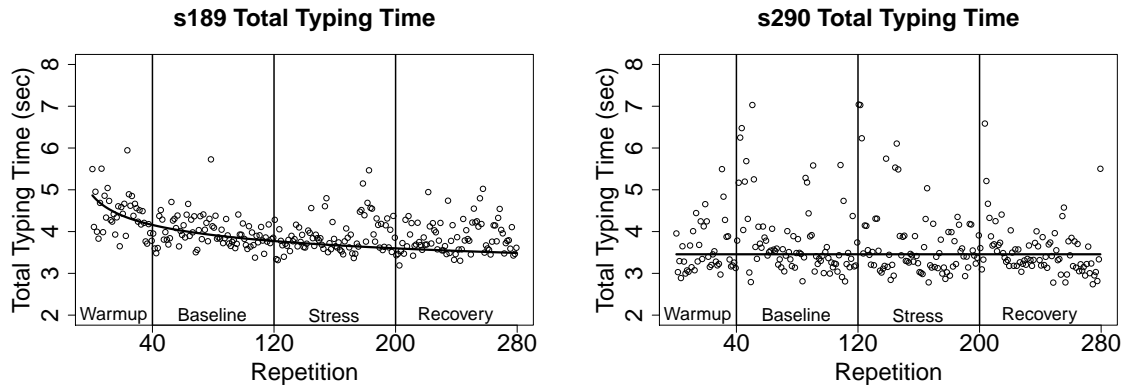
Figure 7.3: **Two total typing time examples.** Two subjects are depicted, with (left, s189) and without (right, s290) a practice effect. In each plot, the total typing time for each repetition is represented by a hollow circle. Vertical bars denote the boundaries between the warmup, baseline, stress, and recovery sessions. A practice curve is also depicted; this curve is fitted using only the data from the warmup and baseline sessions, to avoid potential contamination from the stress typing data. Note that in the left plot, there is a steep decline in the fitted practice curve during the warmup session and, to a lesser extent, in the baseline session. This indicates that there is a noticeable practice effect in the user's typing. In the right plot, the fitted practice curve is effectively a flat line, with no deviation. The flatness of the curve indicates no practice effect is present.

The parameters of each practice curve (M, B, E, and β) are chosen to minimize the sum of the absolute error of the residuals, similar to the process of finding a "best fit line" in linear regression.

Two examples of practice curves are depicted in Figure 7.3. In each example, the total typing time (total time to completely type one full repetition of the phrase) for two subjects, s189 (left) and s290 (right), are plotted against the repetition number. The plot further depicts the four typing sessions that are collected, in chronological order: 1) warmup/familiarization, 2) neutral baseline, 3) stress, and 4) recovery. Overlaid on top of the data are practice curves; these curves are fitted using only the data from the warmup and baseline neutral sessions, since these are the only data known to occur in a non-stressed state. Note that s189 has an evident practice effect in his typing, while s290 does not. This might be explained, for example, by the fact that s290 is a fluent touch typist, while s189 is not.

Hold times -- # subjects practiced by end of warmup session

DD times -- # subjects practiced by end of warmup session

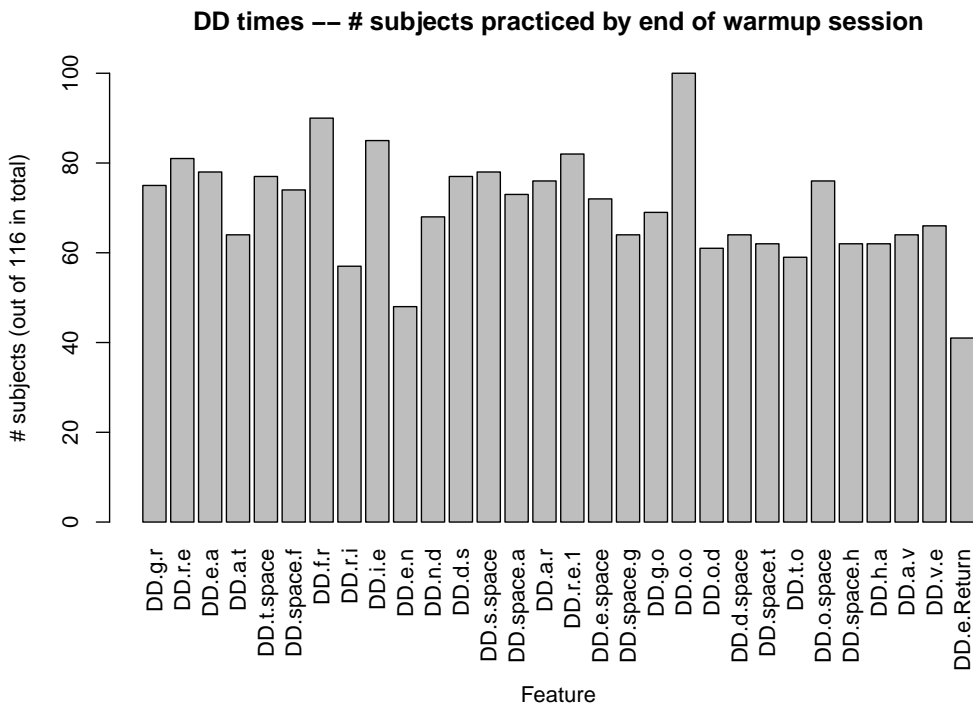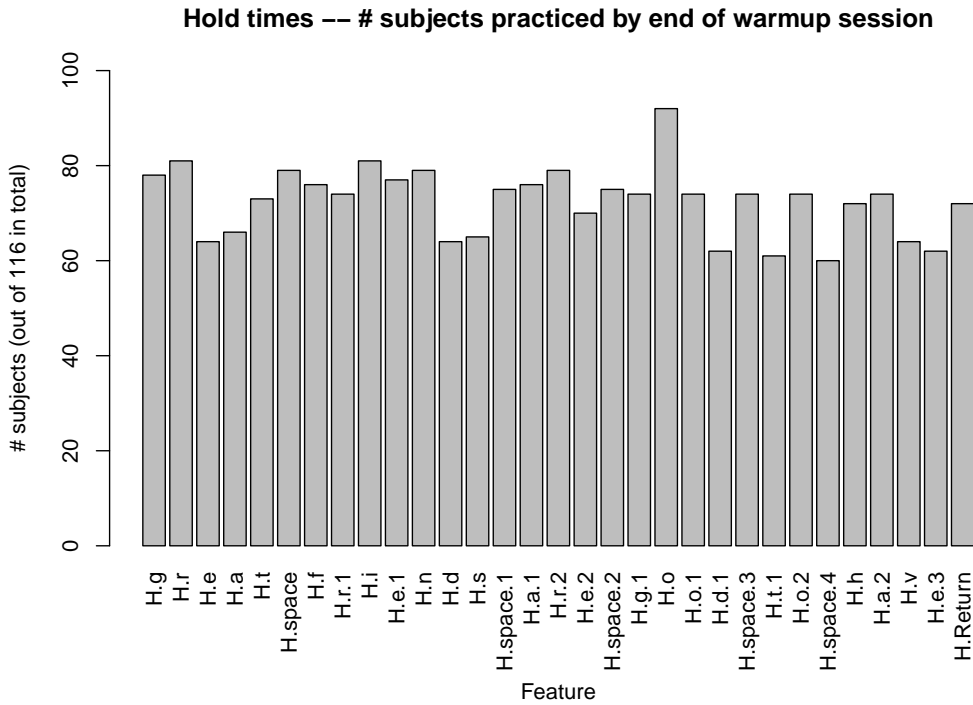Figure 7.4: **Practiced points.** The height of each bar represents the number of subjects that achieve practice, at a 0.1% threshold level, before the end of the warmup period for the hold times (top) and latency times (bottom). Note that a majority of subjects achieve practice before the end of the warmup session for each feature, but that no feature has all subjects achieving practice before that point.

63

7.2.2 Extent of practice effects

Now that we have a method for quantifying practice effects in typing data, the next natural inquiry is the extent to which practice manifests within the data we have collected. We will define a subject as being *practiced* on a given keystroke feature if the expected per-repetition change, as predicted by the fitted practice curve, is less than 0.1%, which is a very conservative threshold. For example, a subject with a keystroke feature with a 1000 ms (1 second) duration would need to have an expected per-repetition change of 1 ms to be considered practiced. We choose to use such a conservative threshold because it guarantees that any typing changes due to practice will be almost nil. Now that we have defined what it means for a subject to be practiced on a given keystroke feature, we define a subject's *practiced point* for a particular feature to be the repetition number after which all expected per-repetition changes, as predicted by a fitted practice curve, are less than 0.1%. Note that the form of a practice curve – a decaying exponential – guarantees that such a point must exist for every feature for every subject.

In an ideal world, all subjects would achieve a practiced state on every single keystroke feature before the end of the warmup session. As we do not use data from the warmup session for analytical purposes, outside of the present examination of practice, any practice effect manifesting solely in the warmup session does not affect any of our other analyses.

Figure 7.4 shows the number of subjects (out of 116) that achieve a practiced state for each hold and DD latency time during the warmup session. With a small handful of exceptions, a majority of subjects become practiced on most features before the end of the warmup period. However, there is no feature for which all subjects become practiced by the end of the warmup period.
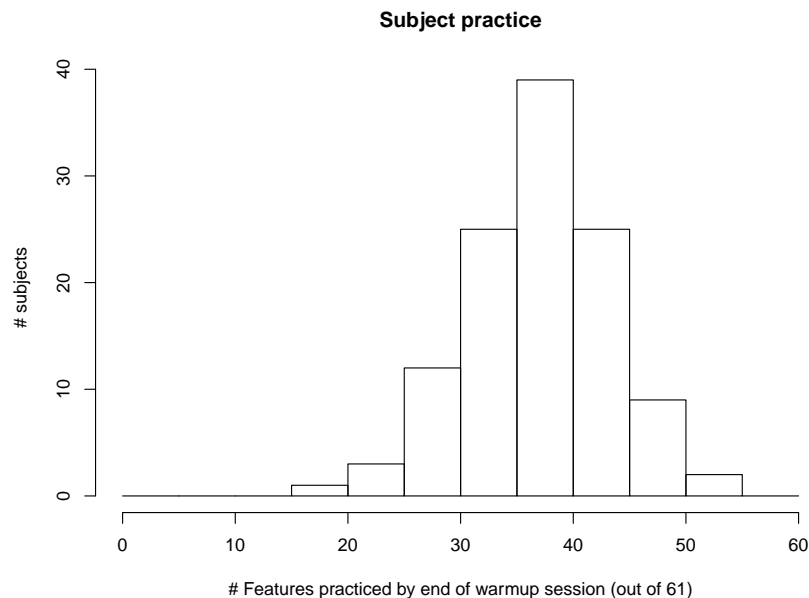

Figure 7.5: **Subjects, number of features practiced by the end of warmup session.** Histogram of the number of features subjects have practiced by the end of the warmup session. A 0.1% threshold for practice is used for all features. Note that a majority of subjects have more than half of their features practiced, but clearly not all subjects are practiced on all features.

64

	AB	BA	ABA	AA			AB	BA	ABA	AA
Sequential	76.64	72.14	71.83	84.87		Sequential	82.47	75.17	72.11	92.25
Random	89.47	80.96	86.04	93.67		Random	92.00	83.46	87.35	97.05

Table 7.4: **Classification accuracies before and after accommodating practice.** Classification accuracies before (left) and after (right) accommodating for practice. In both tables, presented results are averaged over all 116 subjects, with the Random Forest classifier, under all 8 classification regimes when accommodating practice. Note that, as expected, classification improved slightly when accommodating practice. Also as expected, increases were generally larger for the sequential data selection method.

A similar observation holds true when examining individual subjects – most subjects become practiced at a majority of features – but no subject achieves practice on all features. The most commonly practiced feature was DD.o.o (100 subjects), and the least commonly practiced feature was DD.e.Return (48 subjects); this is unsurprising, as repeating a key twice is a common and simple gesture that is likely to be highly practiced while striking the Return key after a letter key is highly uncommon during normal typing and therefore likely to be fairly unpracticed. Figure 7.5 shows the distribution of the number of practiced features per subject by the end of the warmup session. Note that the majority of subjects have at least half of their features practiced.

Regardless of which features or subjects are examined, it is clear that there are practice effects in the collected typing data. Therefore, we now turn our attention toward removing these practice effects from the collected typing data. In doing so, we will confirm whether successful classification is possible after practice effects have been removed from the data.

7.2.3 Accommodating practice effects in typing data

Our technique for removing practice effects from the collected typing data is quite simple. The fundamental principle is that practice manifests in a known manner – in the form of an exponentially decaying practice curve. By fitting a practice curve to each feature for every subject, we can chart out the expected behavior for that feature over time. Then, instead of examining the actual values of the typing data, we instead examine the residuals – the difference between the actual values of the typing data and the practice curve. By doing so, we essentially "subtract out" the effect of practice, leaving only changes in typing from other sources (e.g., stress).

To be more precise, a practice curve is fitted for each feature for every subject using the data collected in the warmup and baseline typing sessions. As with the curves fitted when we were concerned with merely measuring practice effects, our objective here is to use only data that are "untainted" by a stress effect. Once these curves are fit – 61 for each subject – the residuals for every feature are computed by taking the difference between the typing data and the value of the practice curve. The resulting residuals are then used for classification, in place of the original typing data.

Table 7.4 depicts the increases in classification accuracy, averaged over all subjects, for the Random Forest classifier when accommodating practice. All 8 classification regimes are included in the table. Note that accommodating for practice increases classification accuracy under all of the classification regimes, though the increases are relatively minor.

Trend	Avg. number of features
Hold times, expected trend	20.46
Hold times, monotonic trend (not including warmup)	10.54
Hold times, monotonic trend (including warmup)	3.34
Latency times, expected trend	19.55
Latency times, monotonic trend (not including warmup)	10.45
Latency times, monotonic trend (including warmup)	3.24

Table 7.5: **Keystroke feature trends.** Number of keystroke features, averaged across all subjects, following each of the trends found in our data. The expected, return-to-baseline trend involves a shift in the median of a keystroke feature when moving from the neutral baseline to the stress session, followed by a shift in the opposite direction (i.e., a return) when moving from the stress session to the recovery session. Features not displaying this expected trend display a monotonic trend. This monotonic trend may or may not persist when the warmup data are examined.

7.2.4 Conclusion: Practice is not a dominant signal

As we have seen, there is certainly a practice effect in the typing data that we have collected as part of this experiment. While most features for most subjects are, in fact, practiced by the end of the warmup session, it is clear that not all subjects are practiced on all features. However, once the influence of any practice effects has been accommodated by virtue of examining the residuals instead of the original typing data, we can see that successful classification between neutral and stressed typing data is still possible.

Therefore, it is clear that practice is not the dominant signal in the collected data; if it were, then removing the practice effect would significantly lower classification accuracies. Since removing the effect did not do so – in fact, it slightly increased classification accuracy, we are led to conclude that the dominant signal in the collected data is due to stress or something else, but not practice.

7.3 Explaining high AA accuracies

While there is clear evidence of a practice effect within our collected data, it is also clear that this effect is fairly minimal. We have seen that it is fairly implausible that practice alone can explain the high baseline vs. recovery (AA) classification accuracies that we saw in Section 7.1.4. With one plausible explanation ruled out, what might explain these high accuracies?

To answer this question, we must examine the pattern of behavior for individual keystroke features. More specifically, we are interested in whether a given keystroke feature, for a given subject, exhibits *return-to-baseline* behavior. For every keystroke feature, we will observe some change between the median in the neutral baseline typing session and the stress typing session. If this change is an increase, the feature exhibits the return-to-baseline behavior if the median of the recovery session is lower than that of the stress session; if the change is a decrease, a feature exhibits the behavior if the median of the recovery session is higher than the stress session. This is, of course, the expected behavior for any keystroke feature as we expect that typing in the recovery neutral session should begin to revert to the typing of the baseline neutral session when the stressor is no longer present.

If a feature does not exhibit return-to-baseline behavior, we say that the feature exhibits *monotonic* behavior. The median either always increases as we progress chronologically through the

experiment or it always decreases. Features possessing a monotonic trend could be undergoing a practice effect. To determine whether this is the case, we examine the behavior of the feature in the warmup session. If there is a truly a practice effect, the monotonic trend would still be apparent even when taking into account the warmup session.

Table 7.5 shows the average number of features, per subject, that follow each of the identified trends. Note that most of the hold and latency times, approximately two-thirds, exhibit the expected return-to-baseline trend. About one-third of the features show a monotonic trend when not taking into account the warmup session. When the warmup session is taken into account, there are only about three features, on average, still exhibiting the monotonic trend.

A key aspect of ML classifiers, including Random Forest, is that their performance is dependent on the features that are most useful for the classification task at hand. A classifier charged with performing an AA classification task will seek out the one-third of features that possess a monotonic trend, in spite of the fact that a supermajority of the features exhibit the expected return-to-baseline trend and would thus be relatively poor for classification.

Focusing further on these features, we can see that while approximately 3 hold and latency features per subject display a monotonic trend from the warmup session through to the recovery session, there are roughly 7 hold and latency features per subject that display a monotonic trend only within the three non-warmup sessions. This is unexpected behavior. The monotonic trend for these 7 hold and latency features are clearly not caused by practice since the monotonic trend is broken if the warmup session is included. Despite this, not only do these features not exhibit an expected return-to-baseline behavior, they actually deviate further from baseline in the recovery session. It remains an open question as to why this occurs. Our speculation is that subject may have internalized some of the typing changes that were caused by stress. These internalized changes might persist, or indeed strengthen, even when the stressor is removed.

7.4 Statistical search for markers

We have seen in Section 7.1 that we can successfully and reliably distinguish between neutral and stressed typing from the same subject. If we can successfully distinguish the two, we must conclude it is because of some underlying changes in typing. These changes would be precisely the markers that we seek.

We rely on more traditional statistical analyses to reveal these markers, as the models learned by ML algorithms are not guaranteed to be readily human-interpretable. Our approach is simple; we look for features that are significantly different between neutral and stressed typing for each subject.

7.4.1 Identifying markers

We define a marker as any feature whose mean shifts by more than 10% between the neutral baseline and the stress typing sessions. While the choice of 10% as the required shift is somewhat arbitrary, the upcoming analyses can be repeated for any level of shift. We choose 10% as we feel this marks the minimum meaningful mean shift; a shift of this size is unlikely to be caused by random noise. We also choose to compare only the neutral baseline and stress typing sessions, notably omitting the recovery baseline session. As we have previously noted within this chapter, many of our subjects experience an incomplete return to baseline. Accordingly, including the recovery typing when computing the size of the shift would inordinately bias against finding markers.
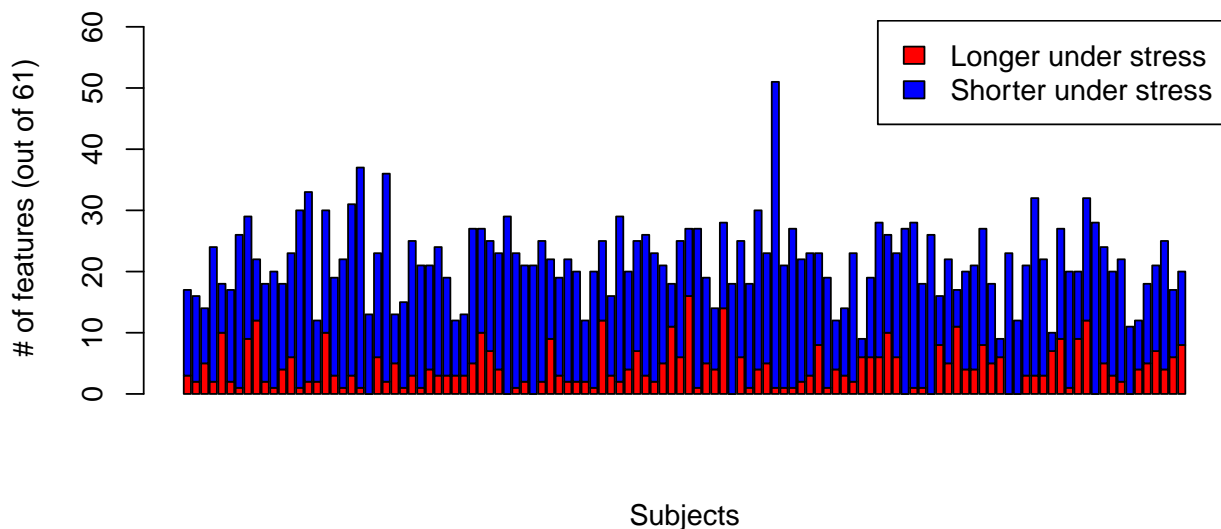
Figure 7.6: **Markers per subject.** Barplot depicting the number of markers for each subject. A feature is considered a marker if its mean differs by more than 10% in the baseline neutral typing session as compared to the stress typing session. Red portions of the bar indicate markers that are longer under stress; blue portions indicate markers that are shorter under stress. Subjects are arranged chronologically from first (left-most bar) to last (right-most bar). Note that all subjects have at least one marker.

7.4.2 Results and Discussion

Figure 7.6 depicts the number of markers for each subject. The height of each bar depicts the total number of markers for that subject, while the colors indicate whether the corresponding feature was longer (red) or shorter (blue) when the subject was stressed. Subjects are organized chronologically from left to right. Critically, note that all subjects have at least one marker; the minimum number of markers was 9 for subjects s264 and s281. Therefore, we have succeeded in our goal in identifying at least one marker per subject. It is also interesting to note that no subject had all features as markers; the maximum number of markers was 51 for subject s253.

7.5 Summary

We have seen in this chapter that we can reliably discriminate between neutral and stressed typing within a single subject. This is true for each of the three classifiers we tried – Random Forest, Support Vector Machine, and Lasso-regularized logistic regression. It is also true in each of the eight classification regimes, which are the cross-product between two data selection methods – sequential and random – and four classification tasks – baseline neutral vs. stress (AB), stress vs. recovery neutral (BA), baseline and recovery neutral vs. stress (ABA), and baseline neutral vs. recovery neutral (AA).

We also saw that there are small practice effects present within the typing data we have col-

lected, but that these do not constitute a dominant signal in our typing data. These practice effects can be removed from our data by fitting a practice curve to each feature for each subject and then classifying the residuals. Doing so slightly increases the accuracies that we obtain.

We also noted that while most features for most subjects exhibit the expected, if incomplete, return-to-baseline trend in the recovery typing session, there are roughly a third of features that do not do so for each subject. The existence of these features suggests that some of the changes in typing that are attributable to stress persist even when the stressor is removed. It remains an open question as to the mechanisms behind this persistence.

Finally, we saw that every single subject has at least nine markers that are substantially different between neutral and stressed typing. Within the next chapter, we will perform a closer examination of the marker patterns across different subjects to ascertain whether there are any universal markers.

Chapter 8

Question 2: Seeking universal markers for stress

We saw in the previous chapter that we are able to successfully discriminate between neutral and stressed typing *within* a single subject. Not only that, we were able to identify at least nine markers for each subject that discriminate their neutral and stressed typing. In this chapter, we wish to perform a more difficult version of both these tasks. We will start by attempting to successfully discriminate between neutral and stressed typing *across* subjects: the objective is to successfully discriminate between neutral and stressed typing from a subject **without** access to data from him/her. If we are able to successfully accomplish this, the potential applications for detecting stress through keystroke dynamics will significantly increase. When performing within-subject classification, we require access to training data for that subject; this means we must have been able to collect neutral and stressed data from that subject ahead of time. This may be realistic in some environments, such as in a secure operating facility where a known set of operators will repeatedly interact with the system. It is not realistic in general environments, where the set of users may be unknown and where subjects may only interact with the system once.

Within this chapter, we will unfortunately see that across-subject classification appears to be elusive. A representative sample of off-the-shelf classifiers was unable to successfully and reliably discriminate between neutral and stressed typing data. Even when using a custom-crafted state-of-the-art deep net, we were unable to obtain above-chance levels of classification.

We believe that the inability to perform this classification is a direct result of strong individual differences in the manifestations of stress in our subjects' typing. We will make the case for this within this chapter by directly examining the lack of pattern(s) in the markers for our subjects.

8.1   Classification

Our approach toward across-subject classification is quite similar to our approach to within-subject classification that we saw in Chapter 7. As we did in that chapter, we will attempt to use the three off-the-shelf classifiers to classify between neutral and stressed typing. The major change is that we will be classifying across subjects instead of within subjects; that is, we will attempt to classify a given subject's typing data as neutral or stressed **without** having seen other data from that subject. We will be relying on data from other subjects, from both neutral and stressed conditions, for training our classifiers.

8.1.1 Classification regime

We utilize a leave-one-out (LOO) classification regime. We will take each subject, in turn, as the test subject (the "left out" subject). All other subjects will be considered training subjects. Thus, our training data will consist of neutral and stressed typing from 115 subjects while our test data will consist of the neutral and stressed typing from a single subject. For the neutral data, we use only typing from the neutral baseline session, omitting the data from the recovery neutral session. As we saw in Chapters 6 and 7, there are strong reasons to believe that subjects did not fully return to baseline in the recovery session, so we maximize our chances of success by only including neutral data from the baseline neutral session. We utilize all available repetitions of neutral baseline and stressed typing data. This totals 115 (subjects) x 80 = 9200 neutral reps and an equal number of stressed reps comprising the training data. For the test data, we again use all available repetitions of neutral baseline and stressed typing data. This gives us 80 repetitions of neutral and stressed data for testing.

The LOO procedure is repeated 116 times, with each subject taken as the "left out" subject once. This produces 116 total accuracies, which are averaged to produce accuracies for each of the employed classifiers.

8.1.2 Classifiers

As in Chapter 7, we will start by employing three readily-available off-the-shelf classifiers: 1) random forest (RF), 2) support-vector machine (SVM), and 3) lasso-regularized logistic regression (LASSO). All details pertaining to the classifiers are unchanged from the analyses performed in Chapter 7, only the classification regime and evaluation procedures change, so we refer the reader to Section 7.1.2 for the full details.

8.1.3 Evaluation procedure

The evaluation procedure employed is straightforward: each classifier is evaluated under the LOO regime and the resulting accuracies are averaged to obtain the classification accuracy for that subject. Thus, for each of the three classifiers, a total of 116 models are trained and evaluated, one for each subject. This produces 116 accuracies, whose average is then the classification accuracy for that classifier.

8.1.4 Results and discussion

Classifier	Avg. Accuracy
RF	59.27
SVM	58.21
Lasso	57.34

Table 8.1: **Across-subject classification accuracies.** Classification accuracies, averaged over all subjects, when training classifiers to discriminate between neutral and stressed typing in a leave-one-out fashion. Note that chance level accuracy is 56.88% (as detailed in the text), so all of our classifiers perform barely above chance.

Table 8.1 depicts the results for classification. Note that the required accuracy for statistically significant classification (i.e., above-chance classification) is 56.88%, corresponding to 91 correct

classifications out of 160 repetitions. This value is obtained by forming a one-sided (greater) binomial 95% confidence interval with 160 total trials (the size of the test set) and a null hypothesis of 50% accuracy. The smallest number of successes (i.e., correct classifications) required for this confidence interval to not include the null hypothesis is the required number of correct classifications to obtain above-chance accuracy. In our particular case, 91 correct classifications are required, corresponding to 56.88% accuracy.

While all of the obtained classifier accuracies are slightly above chance, classification accuracies below 60% are far from convincing. It would be difficult to foresee a practical use for a classifier that performed so poorly in tightly-controlled laboratory settings.

An obvious paradox is why across-subject classification is so poor when within-subject classification is excellent. The most obvious answer is that each individual subject has their own set of markers that make within-subject classification possible, but that these markers are not shared in any meaningful way across all subjects. Without some universal sharing of markers, reliable across-subject classification is impossible for the three classifiers we have employed. Additionally, we are attempting to perform classification using hold and latency times. Even if subjects were to share markers, individual differences between subjects in their natural typing speed would make classification difficult. For example, consider the following scenario. Subject A and Subject B share the hold time on g as a common marker for stress. Subject A has a hold time of 100ms on g during neutral typing, but a hold time of 80ms during stressed typing. Subject B, on the other hand, has a hold time on g of 120ms during neutral typing and 100ms during stressed typing. An across-subject classifier must classify g hold times of 100 ms as being either neutral or stressed typing; in the process, data from either Subject A or B must be misclassified.

Since our off-the-shelf classifiers were unable to effectively perform across-subject classification, a logical step would be to revise the classifier being used. A particularly appealing option, and one that is currently popular in state-of-the-art research and practice, are deep neural networks. Such deep nets have risen in popularity recently with the increase of computing power and are widely employed in a variety of real-world tasks (LeCun et al., 2015). They possess a marked increase, compared to our three off-the-shelf classifiers, in their expressiveness and ability to discover underlying commonalities in stressed typing that are shared across all subjects.

8.2 Deep neural network

The fundamental idea behind our network is the hope that commonalities exist between neutral typing data from all subjects and that they also exist between stressed typing data from all subjects. These commonalities may not necessarily be apparent when examining the original features – i.e., hold and latency times – but may be more obvious once the data have been transformed. As such, we approach our network from the perspective that we must discover an appropriate transformation that permits successful classification.

We begin our discussion of neural networks with a brief overview for readers who may not be familiar with the topic. Then, we discuss the particulars of our network, including the structure, loss function, and training procedure.

8.2.1 Overview

Neural networks have been known since the 1960s, but have recently surged in popularity due to the massive increase in available computational power. A typical example of a standard neural
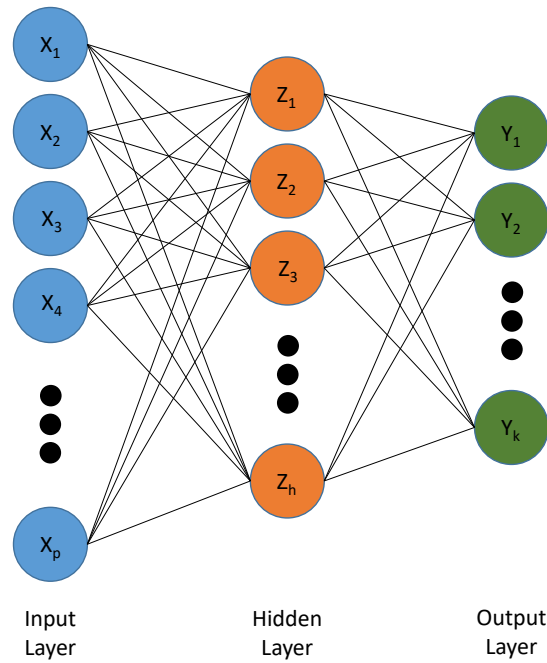
Figure 8.1: **Basic three-layered neural network.** A depiction of a simple, three-layered neural network. Inputs to the network (x_1, ..., x_p) are features (e.g., hold and latency times). The hidden layer of the network (z_1, ..., z_h) consist of multiple nodes. Each node takes as input all of the features x_1 to x_p and computes a score, which is typically a weighted linear sum: $\sum_{i=1}^{p} \alpha_i x_i$, where the αs are weights. This sum is then passed through a non-linear *activation* function. The final layer (y_1, ..., y_k) consist of nodes whose values are a weighted sum of the hidden layers: $\sum_{i=1}^{h} \beta_i z_i$, where the βs are weights.

network is depicted in Figure 8.1. The input layer (blue, leftmost) consists of the features inputted to the network. In the case of the present work, these would be keystroke features (i.e., hold and latency times). The hidden layer (orange, center) consists of nodes whose values are weighted linear sums of the inputs, transformed by some non-linear function (e.g., the logistic function). This non-linear transformation is crucial, as it makes neural networks markedly more expressive than a simple linear combination; without a non-linear transformation function, a neural network of any depth can be equivalently represented by a weighted sum. The outputs of the network (green, rightmost) are typically just a weighted linear combination of the values of the hidden layer.

The parameters of the networks are the weights, both in the hidden layer and in the output layer. By adjusting these weights, one can adjust how the inputs of the network relate to the outputs. In a typical binary classification example, the output layer might consist of 2 nodes, one representing each of the classes (e.g., one node for neutral and one node for stress). The objective would then be to select the weights such that inputted neutral typing would produce a value of 1 in the neutral node and a value of 0 in the stress node while also ensuring that inputted stress typing would produce a value of 0 in the neutral node and a value of 1 in the stress node.

A useful interpretation of the hidden layer is to think of it as feature creation. New features are generated, which are a non-linear function of the original features. The output (i.e., the predicted class of the input) is then a weighted linear combination of these generated features.

The art of training a neural network is a deeply-studied topic. However, all methods share some commonalities. A *loss function* is defined, which represents how poorly the neural network performed. In the above binary classification example, we could choose the loss function as follows. For every repetition, we could take the predicted class to be whichever of the output nodes had higher value. A correct prediction would then incur a loss of 0, while an incorrect prediction would incur a loss of 1. The total loss would then be the sum of the loss for each repetition; this is equivalent to counting the number of misclassifications [1]. Given a set of weights, which define a neural network, the loss can be computed. This information is then used to adjust the weights of the network, typically through a back-propagation algorithm (Hastie et al., 2009, p.395), to better minimize the loss.

8.2.2 Loss function

The primary objective of our neutral network is to transform the data to permit across-subject classification. More precisely, a transformation can be considered successful if it places neutral typing repetitions close to each other and stressed typing repetitions close to each other while also keeping neutral and stressed typing repetitions apart. A popular method for creating such a transformation is to use a triplet approach (Hoffer and Ailon, 2014). In the triplet approach, triplets are formed, consisting of an *anchor*, a *similar* example, and a *different* example. The anchor is a repetition of typing that is either neutral or stressed. The similar example must be of the same class as the anchor (i.e., neutral if the anchor is neutral, stressed if the anchor is stressed). The different example must be of the opposite class of the anchor. When training the neural network, these triplets are fed through the network (i.e., transformed) and a loss is computed for the triplet.

For the moment, we consider a generic distance function d. The loss function is dependent on the distance between the anchor and similar example (d_{sim}) and the distance between the anchor

---

[1]Note that in practice, using this loss function would be problematic for a host of reasons, but we use it as an example since it is easy to grasp.

and the different example (d_{diff}). Intuitively, the loss should be low (or zero) if the anchor is closer to the similar example than to the different example (i.e., $d_{sim} < d_{diff}$); the loss should be high if the reverse is true (i.e., $d_{sim} > d_{diff}$). We define our loss function as:

$$Loss = \sum^{T} max(d_{sim} - d_{diff} - \psi, 0),$$

where T is the total number of triplets prepared and ψ is a small constant that represents the desired margin of correct classification. To understand what the loss function is doing, we can consider two cases.

First, imagine that for a given triplet $d_{sim} - d_{diff} - \psi \leq 0$. In this case, the anchor is closer to the similar example than the different example; moreover, it is closer by at least the margin ψ. The loss for this triplet will be 0, since the desired behavior is obtained.

Second, imagine a triplet where $d_{sim} - d_{diff} - \psi > 0$. In this case, the anchor is either further away from the similar example than the different example or it is closer, but not by the desired margin ψ. In such a scenario, the loss for the triplet will be positive.

Note that the loss is zero if and only if all triplets have the property that the anchor is closer to the similar example than the different example by at least the margin ψ. Since we will be minimizing the loss, we will push the network towards finding a transformation that causes this property to hold for the triplets.

We have so far discussed this loss function in the context of a generic distance function. We wish our network to find a transformation that separates neutral and stressed data from each other while keeping both of the groups similar. A cosine similarity is a natural fit as a metric for this situation. As an added bonus, cosine similarity has fewer issues with degenerate solutions when optimizing, as compared to a metric like Euclidean distance. Metrics like Euclidean distances can often lead a network optimization procedure to exhibit undesired behaviors, like multiplying all weights by a large constant; such a multiplication would naturally increase distances, which may lead to a lower loss function value without improving the transformation learned by the network.

8.2.3 Structure

The network that we employ is a standard neural network with two hidden layers. We vary the size of the hidden layer (denoted R) in our experiments, but for the sake of simplicity we force the two hidden layers to be the same size. The activation function for both the hidden layers is the rectifier function. The rectifier function is the identity if its input is positive and returns zero otherwise; that is, it returns only the positive part of the input. The output layer of the neural network is also of size R and we use a simple weighted linear summation function at the output layer. Note that when R is less than the number of keystroke features, the neural network is implicitly performing dimensionality reduction.

The network can be thought of as following a three-step process. First, in the first hidden layer, new features are created by combining the initial keystroke features. Then, in the second hidden layer, these new features are themselves combined to create a second round of new features. The advantage of having two hidden layers over a single hidden layer that is markedly improves the expressiveness of the neural network. The third step, occurring in the output layer of the network, is that the second round of features are linearly combined to form the new representation of the input data.

R	2	12	22	32	42	52
Test accuracy	51.2%	50.0%	49.5%	50.3%	48.8%	47.9%

Table 8.2: **Neural network classification accuracies.** Test classification accuracies for our custom neural network for various values of R, which is the size of the hidden and output layers of the network. Note that all versions of the classifier performed poorly; in fact, all classification accuracies are at chance levels (i.e., not better than random guessing). The poor classification accuracy is likely attributable to the inherent difficulty of across-subject classification due to strong individual differences.

8.2.4 Implementation and training

Our network is implemented in Python 3.6 (Foundation, 2018) using the Tensorflow architecture, version 1.5 (Abadi et al., 2016). Weights in the network are initialized to small random values, using the Xavier initializer (Glorot and Bengio, 2010). Biases in the network are initialized with the value 1. Training is performed using an Adam Optimizer (Kingma and Ba, 2014), with an initial learning rate of 0.00001. Networks are trained for 100000 iterations, with a batch of 200 triplets used for training at each iteration. Triplets are formed by randomly choosing training data to serve as the anchor, similar, and different examples. The margin is set to $\psi = 0.01$.

The classification itself is performed using a simple k-nearest neighbors algorithm, with $k = 5$, using the implementation found in sklearn, version 0.19.1 (Pedregosa et al., 2011).

We train networks with R (the size of the hidden and output layers) set to 2, 12, 22, 32, 42, and 52. Since we are unsure which value of R will perform the best, we vary R in increments of 10 between the minimum possible size (2) and the dimensionality of the data (61). In accordance with best practices, the data are centered and whitened prior to passing them into the network. Centering sets the mean of each feature to 0 by simply subtracting the mean of each feature from each value. Whitening (Agnan Kessy, 2016) is a transformation that decorrelates the data while also setting all feature variances to 1.

8.2.5 Results and discussion

Table 8.2 shows the testing accuracies obtained by our neural network, for various values of R. The accuracies are, unfortunately, not as desired. The classifier actually underperforms, relative to the three off-the-shelf classifiers that we employed. A superficial explanation for this is that the whitening procedure, which effectively reweights the features, may be downplaying the importance of key features for classification. Ultimately, the root cause for the low classification accuracies is most likely that classification is an inherently difficult or impossible task due to the strong individual differences in stress manifestation for our subjects.

In the remainder of the chapter, we turn our attention to these individual differences. We highlight the extent of these differences and discuss why these differences make across-subject classification so difficult.

8.3 Examining the lack of markers

Thus far in this chapter, we have seen that across-subject classification is unsuccessful despite our successes at within-subject classification. The data used in this experiment are likely to be the best available data, in terms of quality. If the quality of the data is not at issue, what could be responsible

for our inability to perform across-subject classification? As we have suggested within this chapter, and also at the end of the previous chapter (Section 7.4), it is likely that individual differences in manifestations of stress may be responsible for the difficulties in across-subject classification. To address this claim directly, we examine the patterns of markers for our subjects.

8.3.1 Identifying marker patterns

As mentioned in Section 7.4, we define a marker as any feature whose mean changes by more than 10% between the baseline neutral and stress sessions. We saw in that section that there do not appear to be universal markers, but we offered only a shallow analysis at that point. We now perform a more detailed version of that analysis.

We saw in Section 7.4 that every single subject has at least nine markers. A sensible line of inquiry would be to investigate whether there are patterns in the markers for our subjects. Are there one or more marker(s) shared among all subjects? Are there other notable patterns? To examine this question, we use a set of barplot and heatmap visualizations that depict the marker patterns across subjects. We further make the distinction between markers where the feature was shorter under stress and markers where the feature was longer under stress.

8.3.2 Results and discussion

We start by examining how many subjects had each marker as a feature. Figure 8.2 depicts the number of subjects that had each hold time as a marker; Figure 8.3 depicts this for the latency times. Both figures are also color-coded to indicate how many subjects had the corresponding feature grow longer (red) or shorter (blue) when the subject was stressed.

Note that it is immediately obvious that there are no universal markers. A universal marker would have a bar of height 116, indicating that every single subject had that feature as a marker. In fact, the most common markers are the `space-t` and `e-Return` latencies, which are each markers for 69 subjects. Note that this represents less than 60% of our total subject pool. We will provide a deeper-dive analysis regarding universal markers in the next chapter.

Taking a deeper look at the marker patterns across subjects, Tables 8.3 and 8.4 show the marker heatmaps for the hold and latency times, respectively, for our subjects. Within the two heatmaps, features that are not markers are represented by black rectangles, markers that grow shorter with stress by blue rectangles, and markers that grow longer with stress by red rectangles. Subjects (rows of the heatmap) are sorted so that subjects with many markers that were shorter under stress (i.e., more blue rectangles) are near the top of the heatmap while those with many markers that were longer under stress (i.e., more red rectangles) are near the bottom of the heatmap. Features (columns) are in the same order as they occur when typing the phrase. Numbers on the end of a feature name are there to permit discrimination between multiple occurrences of the same feature (e.g., multiple `e`) hold times).

Note that there are no apparent patterns. An obvious marker pattern would involve one or more columns that are entirely or mostly non-black. What we observe, however, is that every column has numerous black entries. Even the features with the most markers have a substantial number, over 40%, of black entries. This means that any purported universal marker misses at least 40% of the population. Moreover, any purported set of universal markers would have the undesirable property that at least 40% of the subjects do not have each marker in the set. With such a significant fraction of the population not being included, it would be difficult to claim that any one marker or set of markers could be deemed universal.
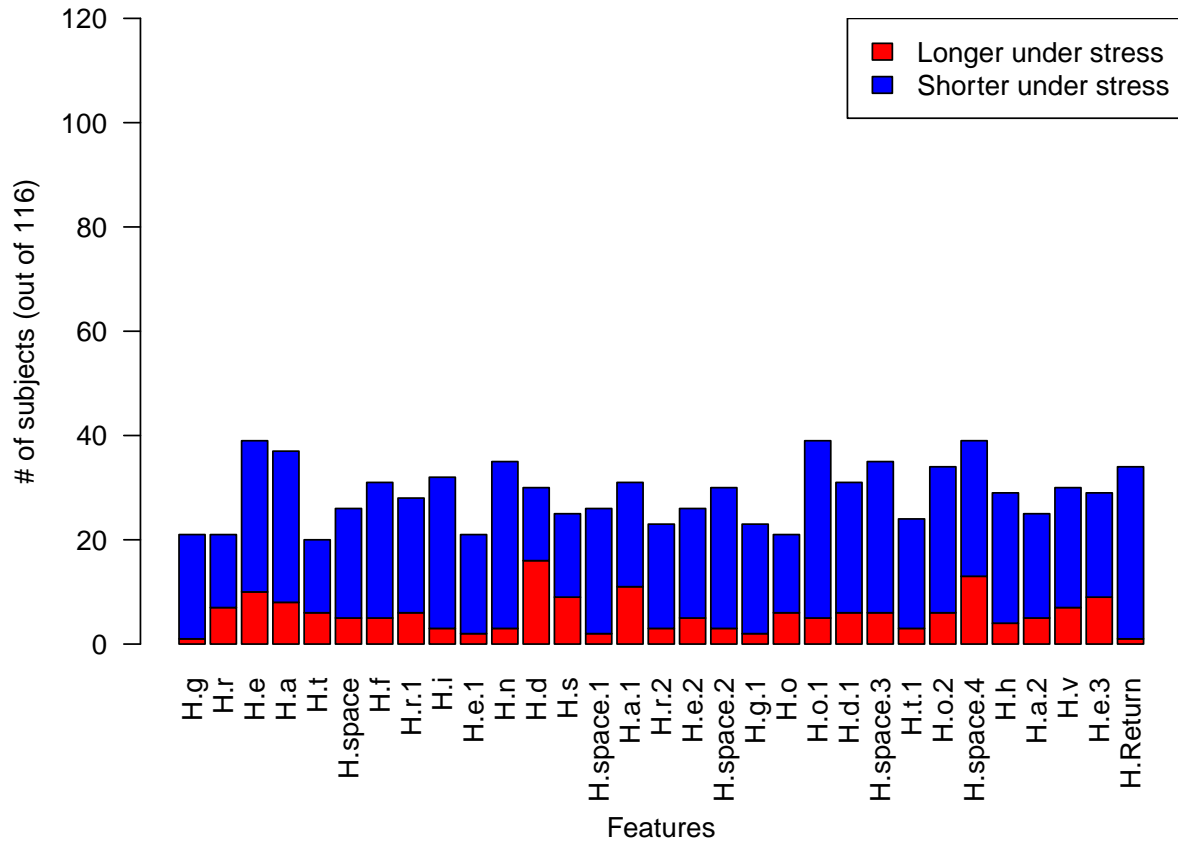
Figure 8.2: **Hold time markers.** Barplot depicting the number of subjects that had each hold time as a marker. To be considered a marker, a feature's mean must differ by more than 10% between the baseline neutral typing session and the stress typing session. Colors indicate whether the corresponding feature was longer (red) or shorter (blue) when the subject was stressed. For example, the hold time for g (leftmost bar) was longer for 1 subject and shorter for 20 subjects. Note that no hold time was a marker for each of our 116 subjects.
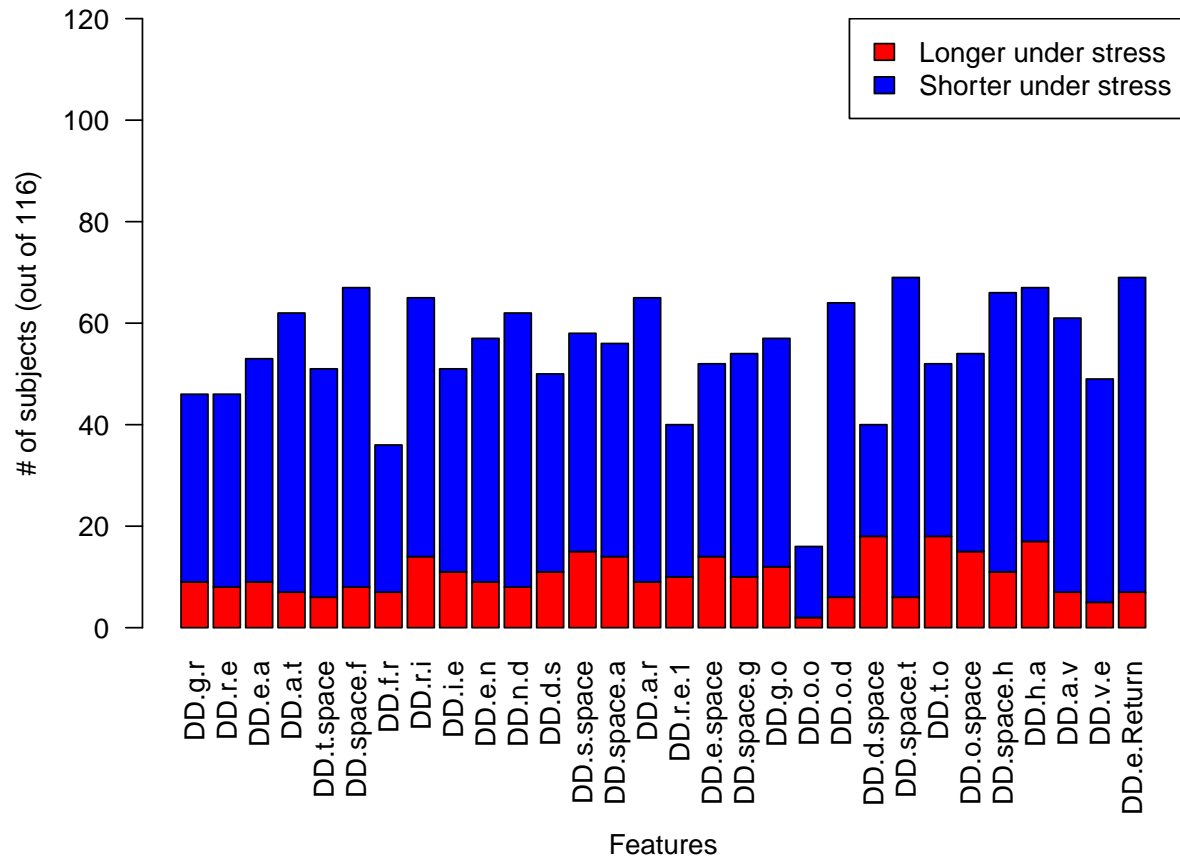
Figure 8.3: **Latency time markers.** Barplot depicting the number of subjects that had each latency time as a marker. To be considered a marker, a feature's mean must differ by more than 10% between the baseline neutral typing session and the stress typing session. Colors indicate whether the corresponding feature was longer (red) or shorter (blue) when the subject was stressed. Note that no latency time was a marker for each of our 116 subjects.

If universal markers, or even near-universal markers, are out of the question, the natural progression of this line of inquiry is to ask whether there may be groups of subjects who have similar markers. As this is a more relaxed definition – different groups could have different sets of markers – we may hope to discover some patterns by searching for and analyzing such groups. The search for such groups is precisely the topic of the next chapter.

8.4   Summary

We have seen in this section that across-subject classification can be performed at slightly above-chance levels using standard off-the-shelf classifiers, but the data apparently do not support sufficiently high classification accuracies to be of any practical use. Moreover, even utilizing a custom deep-net to perform classification did not offer meaningful improvement. Ultimately, the issue appears to be the lack of shared markers for stress among different subjects. Direct examination of the markers indicates that it is fairly obvious that there is no one marker that is universal.

Given that universal markers are not present, a logical fallback would be to inquire whether there might exist groups of subjects with strongly shared sets of markers. This will be the line of inquiry pursued in the next chapter.
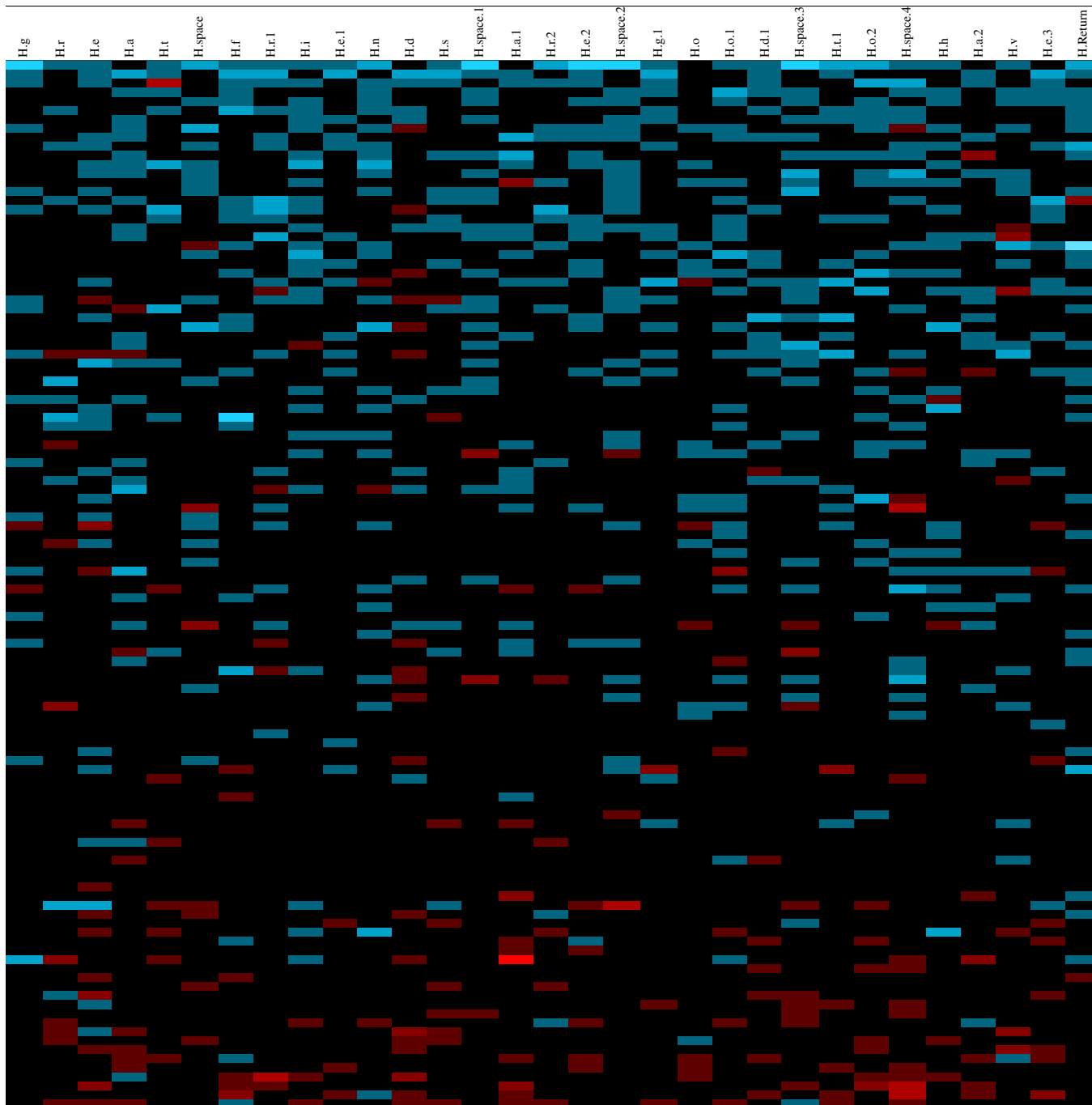
Table 8.3: **Hold time markers.** Hold time markers for each subject are displayed in heatmap format. Black rectangles indicate that a feature is not a marker for a subject. Blue rectangles indicate that a feature is a marker for a subject and that the feature was shorter under stress; red rectangles indicate markers where the feature was longer under stress. More vibrant blue and red rectangles indicate a stronger marker (i.e., larger difference between neutral and stressed). Subjects (rows) are sorted so that subjects with many markers that were shorter under stress are near the top of the heatmap, while those with many markers that were longer under stress are near the bottom of the heatmap. Features (columns) are in the same order as they occur when typing the phrase; numbers on the end of feature names are there to permit discrimination between multiple occurrences of the same feature.
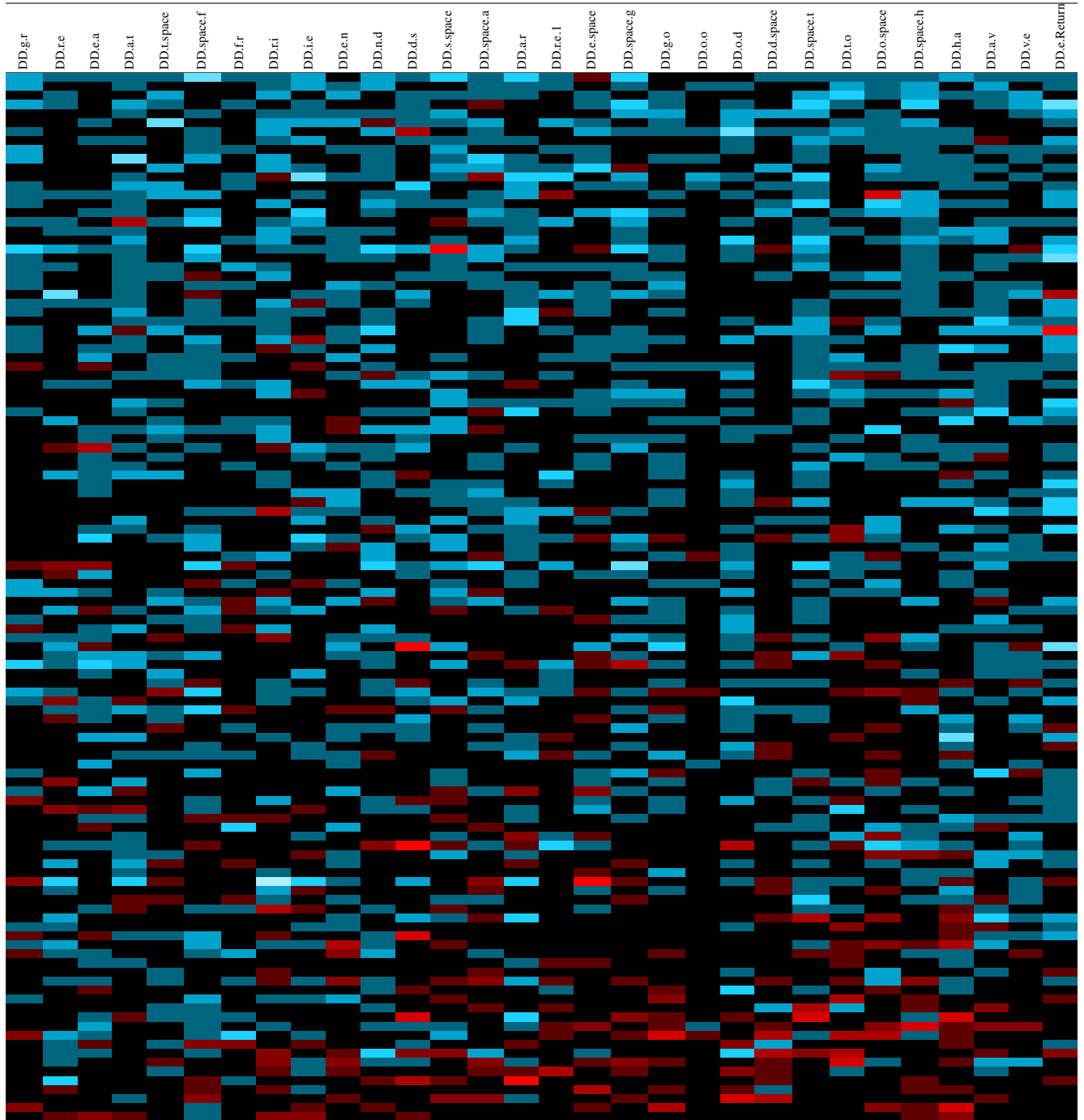
Table 8.4: **Latency time markers.** Latency time markers for each subject are displayed in heatmap format. Black rectangles indicate that a feature is not a marker for a subject. Blue rectangles indicate that a feature is a marker for a subject and that the feature was shorter under stress; red rectangles indicate markers where the feature was longer under stress. More vibrant blue and red rectangles indicate a stronger marker (i.e., larger difference between neutral and stressed). Subjects (rows) are sorted so that subjects with many markers that were shorter under stress are near the top of the heatmap, while those with many markers that were longer under stress are near the bottom of the heatmap. Features (columns) are in the same order as they occur when typing the phrase.

Chapter 9

Question 3: Grouping subjects by response to stress

In Chapter 8, we saw that there are clearly no universal markers. Moreover, there did not appear to be any particular pattern, common to all subjects, to markers for stress in typing. When initially describing our approach to this thesis problem in Chapter 2, we had postulated that there might be a lack of universal markers and offered a contingency plan in that event. This plan, which we focus our attention on now, is to search for groups of subjects with similar responses to stress. The purpose of identifying such groups of subjects is that it may permit classification of stress in unknown subjects (i.e., subjects from whom we have no typing). Such a process would be akin to markers for blood type. If such groups of subjects could be identified with strongly shared markers, we could identify that an unknown subject belonged to a particular group without requiring a priori knowledge of that subject's typing. The fact the unknown subject was a member of a known group could then be used to aid in classifying typing from that subject as neutral or stressed.

As it turns out, searching for groups with commonalities is a fairly well-studied problem. Within the domain of machine learning, this is referred to as *clustering*; the fundamental idea is to find *clusters* (i.e., groups) of data points with common properties. In the case of our specific problem, the objective is to find clusters of subjects (data points) who share commonalities in their markers (properties).

Within this chapter, we will provide a brief overview of clustering, explain how we set up the clustering problem for our particular data, define the metric used to evaluate whether we have successfully found clusters, and present the results of our evaluation. We shall see that, despite our best efforts, there do not appear to be tight clusters of subjects within our data with strong shared sets of markers. This suggests that stress, as manifested in keystroke typing behavior, is a strongly individualized response.

9.1 Clustering

At its most basic, the principle is of clustering is to take a pool of data points and divide them into clusters, such that the data points within each cluster are similar to each other while data points in different clusters are dissimilar. The vast majority of clustering algorithms, including all of those we will entertain in this chapter, rely on a definition of similarity between two data points. A simple example of a similarity measure is Euclidean distance; two points are similar if they are

close together and they are dissimilar if they are far apart. Given a similarity measure, a clustering algorithm then generates a *partition* of the data set into some number of clusters.

Despite this relatively simple sounding description, the actual use of clustering algorithms is often as much an art as it is a science. In some machine learning tasks, like classification, is it straightforward to ascertain the goodness of an outcome by examining a metric like classification accuracy. If a classifier regularly succeeds in predicting the label of test data points, it is a good classifier. Judging the goodness of a given partition from a clustering algorithm is far trickier. Is a result with many small clusters preferable to a result with a few large clusters? Is it acceptable for a result to generate clusters that are winding and elongated (i.e., amoeba-shaped) instead of spheroid? The answers to questions like these are in the eye of the beholder; different observers may differ greatly in opinion on the goodness of a given partition.

An example of a common difficulty in clustering is to simply determine how many clusters one should seek in the data. Clustering algorithms generally require the number of desired clusters as input. The obvious difficulty in providing that number is that we have no idea how many clusters we expect to have in our data. An imperfect solution, and the one that we use in this work, is to define a range of the number of possible clusters and then re-run the algorithm for each value in the range. Such an approach can be successful in the case where the data are naturally strongly partitioned.

Our objective in this chapter is primarily to determine whether there is any evidence of strong partitioning in the stress response of our subjects. A successful outcome would be one in which subjects in the same cluster strongly share some core set of markers. Some subjects within the cluster might have a small number of additional markers, while others might be missing a small number of the core set of markers, but these deviations from the core set should be small.

9.1.1 Clustering setup

At a first glance, clustering our subjects seems quite straightforward. We would pick some set of clustering algorithms to run, feed them our collected neutral and stressed typing data, and then collect the resulting partitions on the other end. There would be, of course, the issue of analyzing the goodness of those partitions, but the clustering itself might seem straightforward. Unfortunately, this turns out to not be the case. The primary difficulty with this naive approach is that our objective is to cluster subjects by their stress response and not to cluster the underlying data.

For example, consider a situation where we performed clustering using this naive approach. Suppose that for a given subject, some repetitions of neutral typing data belong to cluster A, the remaining repetitions of neutral typing data belong to cluster B, and all repetitions of stressed typing data belong to cluster C. How would we identify what cluster this subject belonged to? That is a difficult question to answer, even in this relatively simple example. In a messier example, where a subject's data might be spread across a dozen different clusters, identifying the cluster that a subject belongs to would be even more challenging.

To avoid this thorny issue, our approach is to simply represent each subject as a single data point. Since each subject will be his/her own data point, identifying the cluster that subject belongs to will be trivial; it will simply be the cluster that contains the singular data point. We consider two methods for constructing such a representation. Both methods share the property that they capture a difference between neutral and stressed typing. As with our other analyses, we restrict ourselves to using the only the neutral baseline data to represent neutral typing; the recovery data are omitted since we have seen evidence that subjects did not fully return to baseline.

In the first method (percentage method), we represent each subject by the percent changes in the mean of each hold and latency time between the neutral baseline and stress typing sessions. That is, for each subject, we compute the mean of each hold and latency time in both the neutral baseline and stressed typing session. We then compute the percentage change between those means. Our convention is to have a positive change indicate a longer duration in the stressed typing session; in practice, using the opposite convention should make no difference in the resulting partitions.

In the second method (raw method), we represent each subject by the raw change (in seconds) in the mean of each hold and latency time between the neutral baseline and stress typing sessions. As with the percentage method, we compute the mean of each hold and latency time in the neutral baseline and stressed typing session. However, instead of computing a percentage change, we simply take a difference of means. As with the percentage method, our convention is to have a positive change indicate a longer duration in the stressed typing session, though using the opposite convention should make no difference.

Note that both methods represent each subject by the difference in their typing between neutral and stressed conditions. Thus, when a partitioning places subjects together into a cluster, it indicates that the two subjects have similar changes in typing when stressed. Our goal in this chapter is to determine whether there exists a partition of our subjects that produces tight clusters with strong commonalities among the subjects in that cluster. Such tight clusters would indicate that there are groups of subjects with strongly shared markers. It is worth noting that our clustering algorithms may find several, loosely-packed clusters, perhaps corresponding to high, medium, and low responders to stress or to subjects that typed slower vs. faster when stressed. Such loosely-packed clusters are not of particular interest, as they cannot be leveraged for practical purposes in the manner that tight clusters could be.

9.1.2 Clustering metrics

We have so far seen how we will summarize each subject by a single data point prior to applying a clustering algorithm. While it may seem that the natural next step would be to actually define and apply some clustering algorithms and obtain some partitions, we first turn our attention to metrics for measuring the goodness of a given partition. After all, obtaining a partition is meaningless if we cannot evaluate how good it is.

There are numerous metrics for evaluating the goodness of a given partition; Wang et al. (2009) provides an overview of some of the more popular metrics. Metrics can be broken down into two categories. External metrics compute how well a given partition agrees with some external labelling of the data. For example, if we already knew the clusters subjects should fall into (i.e., we had a "correct" partitioning), an external metric would compare the degree to which a given algorithm's partitioning agreed with the known clusters. Internal metrics do not require such an external label. Instead they attempt to compute the extent to which a given partitioning has clusters that are compact, cohesive, dense, and distinct. As we do not have any external labels, we restrict our attention to internal metrics.

As noted in Wang et al. (2009), common internal metrics include the Silhouette index, Davies-Bouldin index, Calinski-Harabasz index, Dunn index, and RMSSTD index. These metrics all attempt to capture the compactness, cohesiveness, density, and distinctness of the clusters in a given partitioning. Where the metrics differ is in how these terms should be mathematically defined. These differences are akin to the manner in which the concept of spread can be measured using range, inter-quartile range, variance, or standard deviation. The common goal of these metrics is

to measure the underlying concept of spread in a data set, but the metrics differ in their precise mathematical definitions.

For the purposes of this chapter, we will focus on using Silhouette index as our singular metric. We make this choice for several reasons. First, we focus on a single metric for the sake of simplicity. Second, the intuition and mathematics of the Silhouette index are easy to understand. Third, our results are semantically identical when using other internal metrics, and we feel that a full presentation of results, with an array of metrics, would be more difficult to grasp while adding no value.

Computing the Silhouette index is fairly straightforward. Suppose that we have data points $x_1, ..., x_n$. Let $a(i)$ be the average distance x_i to all other data points in the same cluster. $a(i)$ can be interpreted as how well x_i fits in its assigned cluster. If the value is small, x_i is close to every other data point in the cluster; if it is large, at least some data points in the cluster are far away. We can also define the distance between x_i and a given cluster C as the average distance between x_i and every point in C. Let B denote the cluster, not containing x_i, that is the closest to x_i; B can be interpreted as the neighboring cluster to x_i. We can then let $b(i)$ denote the average distance between x_i and the points in B. Then, we can define the silhouette for x_i as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}.$$

Note that the silhouette for x_i tends to be close to 1 if and only if x_i is relatively close to the other points in its cluster while being far away from the points in the neighboring cluster. The silhouette is 0 when $a(i) = b(i)$, which occurs when x_i is equally far away from points in its cluster and the points in the neighboring cluster; this situation can occur if a single cluster were to be randomly split in half. The silhouette is negative when x_i is actually closer, on average, to the points in the neighboring cluster than in its own cluster. A negative silhouette suggests that x_i may not truly belong in its assigned cluster and/or that it might be an outlier in the data set.

The Silhouette index is then defined as the average silhouette for all points in the data set. Intuitively, if this index is near 1, most data points are in tightly-packed clusters and are distant from the neighboring cluster. If the index is near 0, data points generally fit as well in their assigned cluster as with their neighboring cluster. This would suggest that the given partition is poor; if a set of partitions from a variety of clustering algorithms have a low Silhouette index, this would then suggest that there is not a strong underlying cluster structure.

It can be helpful to calibrate one's intuition of the Silhouette metric using a few simple examples. Figure 9.1 contains four such examples with randomly generated data. These randomly generated data sets each have 116 data points, just like our actual data sets. In the upper-left, data were generated uniformly at random within the (0,10) square. The remaining three examples are mixtures of Gaussians with varying degrees of spread. In each, the Gaussians are centered at (3,3), (3,7), (7,3), and (7,7). In the upper-right, the covariance matrix was the identity. In the bottom-left, the covariance matrix was 0.5 times the identity. In the bottom-right, the covariance matrix was 0.05 times the identity. Color-coding of the data points indicate the clusters in each example.

It is worth noting that a Silhouette index is affected by the dimensionality. In the examples depicted in Figure 9.1, the dimensionality of the data is clearly 2. It is trivial, at least mathematically, to expand these examples to higher dimensions. For the uniform example, this can be done by sampling points uniformly at random within the (0,10) hypercube. For the Gaussian examples, we

keep the first two coordinates of the centers the same as in the 2-d example, and set all remaining coordinates to 0. The caption of Figure 9.1 contains the Silhouette index for each example in 2, 5, and 61 dimensions. The value for the 5-d analogues of these examples will be useful later on for a clustering algorithm that utilizes dimensionality reduction; the value for the 61-d analogue is useful for the algorithms that cluster in the original feature space (of hold and latency times).

9.1.3   Clustering algorithms

We use three different clustering algorithms in our work, each representing a different class of clustering methods. The first algorithm is PAM (partitioning around mediods), which represents the most common types of clustering algorithms, which attempt to assign each data point into a single cluster, using the originally provided features. The second algorithm is Agnes (agglomerative clustering), which represents hierarchical clustering methods, which assign each data point into successively broader clusters to create a taxonomy of the data. The third algorithm is the use of LLE (locally linear embedding) followed by a standard clustering algorithm, representing the class of spectral clustering algorithms. The approach of spectral clustering algorithms is to express the data as a graph of its nearest neighbors and then to reduce the dimensionality of the data by expressing data points as a function of its neighbors. Such methods are particularly powerful if the data tend to lie in a low-dimensional subspace of the original feature space.

PAM (Partitioning around mediods). The PAM algorithm is a variant of the well-known k-means algorithm. The PAM algorithm is quite straightforward. To initialize the algorithm, k random data points are designated as mediods (cluster centers), where k is the number of desired clusters. Then, all data points are associated with the closest mediod, as measured by Euclidean distance. The cost of any particular choice of mediods is the summed Euclidean distance between each data point and its assigned mediod. The actual algorithm itself has a single, repeated step. For every pair of mediod and non-mediod points, consider swapping their roles – that is, replace the mediod point with the non-mediod point. If this results in a lower cost, keep the swap. Otherwise, revert the swap. In our work, we use the implementation of PAM within the `R` package `cluster` (Maechler et al., 2017).

Agnes (Agglomerative clustering). The Agnes algorithm is a sub-class of hierarchical clustering techniques, which generate a taxonomy of the data. The fundamental idea behind the algorithm is to take a bottom-up clustering approach. The algorithm is initialized with each data point in its own cluster. Then, at each iteration of the algorithm, the two clusters that are most similar are merged together. This process is repeated until a single mega-cluster, containing all of the data, is formed. By tracing the path of each data point as it gets merged into various clusters, it is possible to create a dendrogram like the one in Figure 9.2. Similarity within the algorithm is defined by the average of the Euclidean distance between each unique pair of points in the two clusters. For example, in the trivial case where each cluster contains only a single point, the similarity between the two clusters is just the Euclidean distance between those two points. If we have a cluster of 3 points and a cluster of 5 points, 3×5 distances must be computed and their average is the distance between the two clusters. In our work, we use the implementation of the Agnes algorithm within the `R` package `cluster`.

LLE (Locally Linear Embedding). The LLE approach to cluster is a two-step process. First, the data are reduced in dimensionality by using a Local Linear Embedding. Then, the reduced-dimensionality data (*embedded data*) are clustered using the aforementioned PAM algorithm. A full description of the mathematics behind LLEs can be found in Saul and Roweis (2000). The fun-
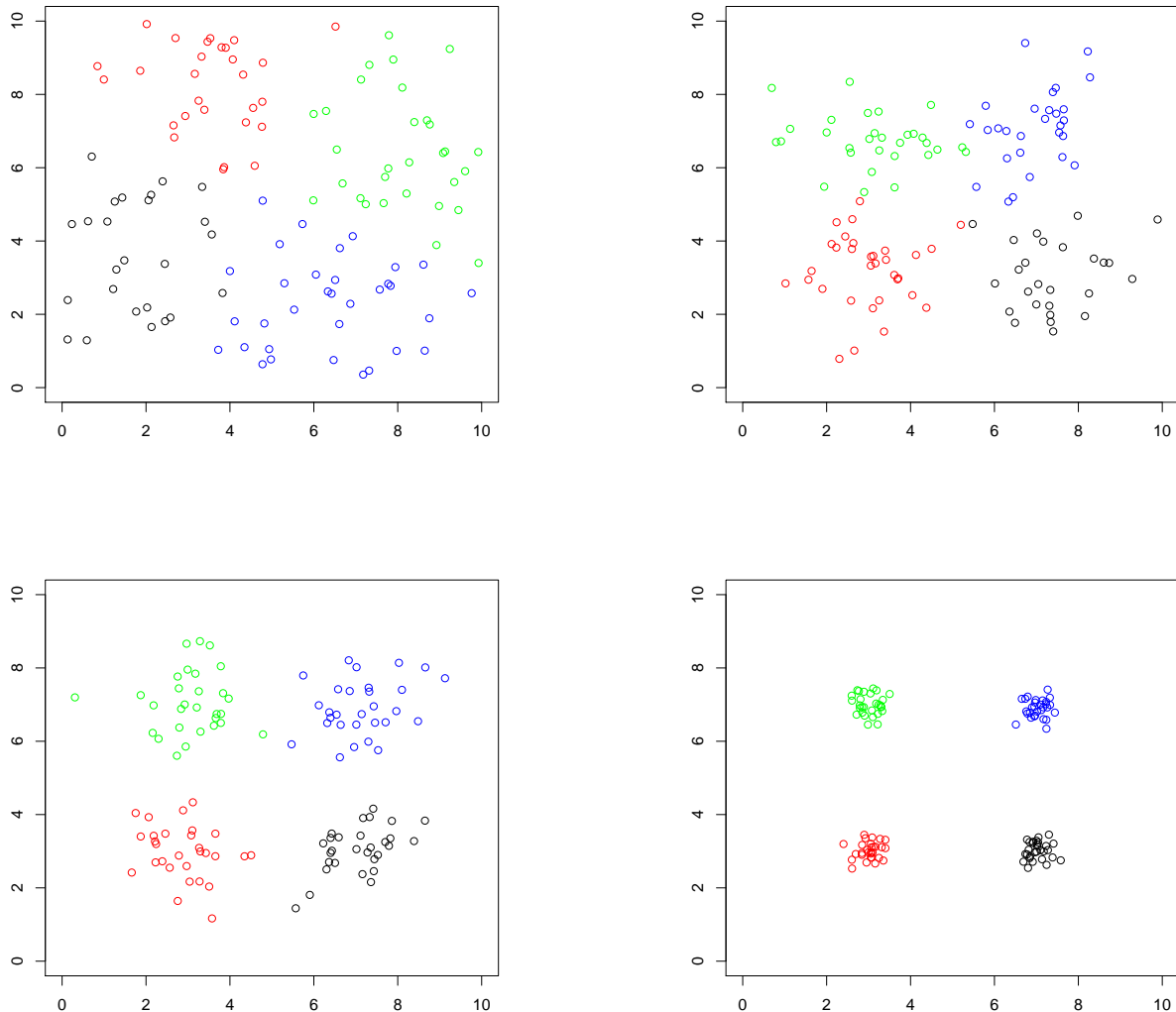
Figure 9.1: **Silhouette Index examples.** Four sample data sets with a partition (indicated by differently colored points) are depicted here. The upper-left plot contains data generated uniformly at random. The remaining plots contain data that are a mixture of four Gaussians with high (top-right), medium (bottom-left), and low (bottom-right) spread. In the two-dimensional examples plotted here, the Silhouette Index is 0.51, 0.46, 0.62, and 0.89 for the uniform, high spread, medium spread, and low spread examples, respectively. The plotted examples also have higher-dimensional analogues (see text for details). The 5-dimensional version of these examples leads to Silhouette Indicies of 0.15, 0.34, 0.47, and 0.82. The 61-dimensional version leads to Silhouette Indicies of 0.003, 0.015, 0.061, and 0.467. We provide these values as context for results presented in this chapter.
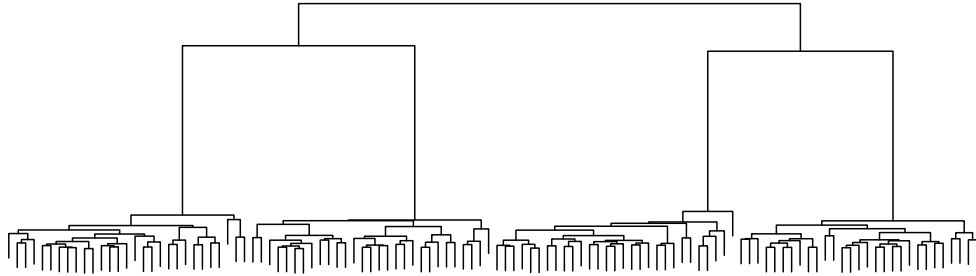
Figure 9.2: **Agnes dendrogram.** A sample dendrogram plot depicting the results from running the Agnes algorithm on the most-tightly clustered example depicted in Figure 9.1. At the top-most level of the dendrogram (the root), all of the data are represented. As one descends through a dendrogram, following successive splits, smaller and smaller portions of data are represented. At the bottom-most portion of the dendrogram (the leaves), a single datum is represented. The length of each arm is proportional to the difference between the data represented in the left and right branch of each split. In this example, the longest arms correspond to the split between individual clusters. The second-longest arms (at the top) represent the difference between the two clusters on the left side and the two clusters on the right side.

damental idea behind this dimensionality-reduction approach is to assume that the data lie within some low-dimensional manifold; this manifold could be simple (e.g., a hyperplane) or highly non-linear (e.g., a high-dimensional swiss roll). Regardless of the complexity of the manifold, it is assumed that the manifold around each data point is locally flat – that is, locally linear. Each data point is then approximated as a weighted combination of its m closest neighbors. These weights then form a new coordinate system for the data. Effectively, this technique attempts to flatten the non-linear manifold, akin to the way a 3-dimensional crumpled piece of paper can be flatted into a 2-d surface.

9.1.4 Clustering algorithm evaluation

Our evaluation procedure for our clustering algorithm evaluation is straightforward. We run each of the three clustering algorithms on both clustering setup methods (raw and percentage). As previously mentioned, each of the three clustering algorithms we use requires us to input the desired number of clusters to find. Since we do not have a firm idea as to how many clusters there will be, we run each algorithm multiple times, each with a different number of desired clusters. Specifically, we vary the number of desired clusters for all values between 2 and 20, inclusive. This range is sufficiently wide to capture any meaningful cluster structure within our data. With 20 clusters, each cluster would contain less than 6 subjects, on average; it is not clear that we could make meaningful statements about clusters smaller than that. After each run of the clustering algorithm, we compute the Silhouette Index for the resulting partition.

Figure 9.3: **PAM Silhouette Index.** Silhouette index for the PAM algorithm's partitions for 2-20 clusters. There appears to be, at best, a loose clustering of subjects, and not the desired tight clustering that would indicate strongly shared markers between a group of subjects.
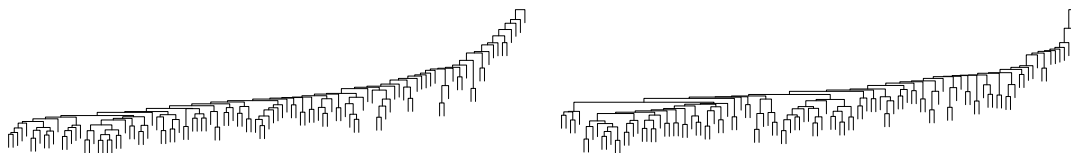
Figure 9.4: **Agnes dendrograms.** Dendrogram plots depicting the results from running the Agnes algorithm on the raw (left) and percent (right) data. Note that in both dendrograms there are no arms that are markedly longer than the average arm length. This suggests that there are no natural clusters in the data we are clustering. Compare these dendrograms against Figure 9.2, where there are clear clusters within the data.

Figure 9.5: **LLE Silhouette Index.** Silhouette index for the PAM algorithm's partitions after transforming the data with a LLE. There appears to be, at best, a loose clustering of subjects, and not the desired tight clustering that would indicate strongly shared markers between a group of subjects.
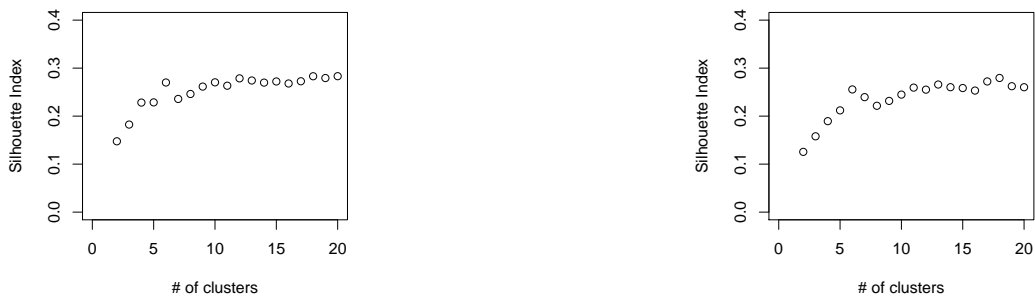
90

9.1.5 Clustering results and discussion

We examine the results for each algorithm, in turn. To place these results in context, we suggest the reader view these results in the context provided by Figures 9.1 and 9.2.

PAM. Figure 9.3 depicts the Silhouette Index for the PAM algorithm, varying across 2 to 20 clusters. Note that the Silhouette Index is roughly 0.02 for most of the cluster sizes for both the raw and percentage data. Comparing to the baseline seen in Figure 9.1, we can see that these values correspond largely to a highly spread case (upper-right example). Direct examination of the few clusterings that produce slightly higher, but still low, values can be seen as separating the data into low, medium, and high responders. These are not the tight clusters that we seek.

Agnes. Figure 9.4 shows two dendrograms resulting from running the Agnes algorithm on the raw and percentage data. It is evident that there is no obvious clustering structure; compare Figure 9.4 to the sample dendrogram in Figure 9.2 generated from data with obvious clusters. The long arms, indicating clear separation between clusters, are simply not present in the two dendrograms generated using the actual data.

LLE. Figure 9.5 depicts the Silhouette Index for the LLE algorithm, varying across 2 to 20 clusters. Recall that the LLE method actually projects the data down to 5 dimensions before using the PAM algorithm to cluster the results. Accordingly, it is important to use the 5-dimensional index values in Figure 9.1. As with the PAM algorithm without dimensionality reduction, the produced partitions for both the raw and percentage data do not generate strong clusters, which would be indicated by a Silhouette Index above 0.8.

9.2 Summary

We have seen in this chapter that little evidence exists for tight groups of subjects with strongly shared commonalities in stress response. Out of the three clustering algorithms that we used, with one representative each from the three major types of extant clustering algorithms, none were able to identify such tight groups. This is true regardless of whether we used raw typing changes or percentage-based typing changes as the input data to the clustering algorithms.

We are therefore forced to conclude that such groups of subjects are unlikely to exist. If they do exist, these groups may be sufficiently small percentages of the population that they are not visible given the sample size of our study. However, we speculate that our inability to detect any such groups is due to the tremendous individual differences in the stress responses of our subjects. It is well-established that natural typing rhythms differ significantly between subjects; it appears that changes in typing rhythm due to stress also differ significantly between subjects.

Chapter 10

Discussion

10.1   Overall findings

The findings of this thesis are divided in terms of the initial questions posed in Chapter 2.

Problem 0: Induce stress in experiment participants.
Findings: We found that our attempt at inducing stress in our subjects, through a combination of a multi-tasking framework and social evaluation, was successful for 100% of our subjects. This was ascertained through statistical examination of our independent-assessment data, consisting of a mix of physiological and psychological evaluations of our subjects. We did note that our subjects' recovery was only partial during the 15-minute recovery period by both physiological and psychological measures (see Figure 6.1). As previously noted, we had conducted our pilot studies with a 30-minute recovery period, but found that subjects became anxious and irritated due to the length of the period, preventing effective recovery. It remains unclear how this paradox can be avoided. Notwithstanding the issues surrounding the recovery period, we were pleased to find strong evidence that our stress induction technique was effective for all subjects.

Problem 1: Characterize how an individual subject's typing rhythms are affected by stress.
Findings: We were able to identify specific markers for stress within each subject. In fact, we were able to identify at least nine such markers for each of our 116 subjects where the marker shifted by at least 10% between the baseline and stress session. Given the existence of these markers, we were also able to successfully classify neutral and stressed data within a given subject. These classification accuracies were well above chance rates, with an average accuracy of 89.5% in the most favorable classification regime. Classification accuracies in this regime ranged from 75.0% to 99.1%; 61 of our 116 subjects (over 50%) had classification accuracies above 90%. Such high accuracies indicate that stress clearly manifests in the typing rhythms of subjects. As our work was intended as a proof-of-concept, we have largely employed simple, off-the-shelf classifiers. We imagine that with refinement of existing techniques and development of new techniques, it would be possible to reliably obtain 90+% classification accuracies for virtually all subjects. At such classification accuracies, it may be feasible to consider pilot testing this technology in real-world scenarios.

Problem 2: Identify universal markers for stress.
Findings: Despite obtaining high accuracies on within-subject classification, we were unable to find universal markers for stress that were common to all subjects. Of the classifiers that we examined, including a deep neural network, the best accuracy was below 60%. While this was still statistically significantly above chance levels of classification, it may be too low to be useful in a practical application. Direct examination of marker patterns helped to reveal why across-subject classification performed so poorly: the most common marker is only shared by 69 out of 116 subjects. There simply are not universal markers. The consequence of these results is that we cannot expect to deploy a system utilizing keystroke dynamics for stress detection in an open-world environment, where the potential users are unknown. Any useful systems must be personalized for each user. This still leaves a large number of applications in closed-world environments, where the list of all users is known; such environments would include air traffic control centers, secure operating facilities, process-control operations, or common office workplaces.

Problem 3: Identify groups of subjects that share common markers.
Findings: We examined our data for evidence of groups (i.e., clusters) of subjects with strongly shared markers, analogous to the protein markers for blood type. Despite trying a variety of clustering algorithms, we found little evidence that such clusters exist. Rather, it seems that subjects have strongly individualized manifestations of stress in their typing.

Overall: At the outset of the experiment, we expected that our stressor would be successful for the vast majority of our subjects. We had strong expectations of discovering markers for stress for individual subjects and felt it was a plausible notion that we would discover either universal markers for stress or groups of subjects with strongly shared sets of markers. Our stressor was effective on all of the 116 subjects run in the study and we were able to identify at least nine markers for each of these subjects. We did not find either universal markers for stressor nor did we find groups of subjects with strongly shared markers. The available evidence suggests that the manifestations of stress in typing data are highly individualized.

10.2   Contributions of this work

Our work makes several major contributions to the existing literature. First and foremost, we view the experimental methodology and protocol of our study as a substantial contribution. All of the previous work in the literature has had significant flaws, ranging from the use of unvetted stressors (i.e., stressors that have not been previously demonstrated to induce stress), to the lack of objective and independent measurements to confirm affect induction, to the presence of multiple uncontrolled confounding variables. Our view is that conducting experiments with stimuli that may or may not actually induce stress in subjects, without actually confirming whether the supposed stressor has had the expected effects on a subject, while having other confounding variables that may be partially or wholly responsible for the obtained results is not a sound mechanism for making scientific progress. In view of numerous works with one or more of these significant flaws, we hope that offering a concrete and detailed experimental methodology and protocol will set an example for other researchers in future work. Moreover, unlike most of the existing literature, we feel that we have offered sufficient detail within this thesis to permit reproduction and replication of this work by others; it is our hope that others will take on this task to support or improve on our

obtained results.

A second contribution is that we have collected multifaceted and detailed data from 116 subjects, which is the largest (and most competently-collected) data set we know of in the literature. Our data include physiological and psychological data to independently confirm the affective states of our subjects, 280 repetitions of typing data from each of our subjects, and numerous supporting data that may reveal relationships between subject attributes and their changes in typing from stress. It is our intent to make all of our data publicly accessible as soon as possible, in the hopes that other researchers may benefit from their availability.

From an analytical perspective, our contribution is that we have performed a deeper analysis than is currently common in the literature. In addition to providing surface-level numbers like classification accuracies, we also investigate the reasons why we obtain the classification accuracies that we do. In the case of within-subject classification, we noted that success could be attributed to the presence of at least nine markers for each subject. In the case of across-subject classification, we noted that our poor classification accuracies could be directly attributed to the lack of universal or near-universal markers. This deeper analysis provides insight into why we obtain the classification results that we did.

A further contribution of this work is that we established a protocol for choosing typing stimulus given a set of criteria. While this did not receive much focus within the scope of the thesis, with many of the details left to the Appendix, this is a common problem in keystroke dynamics research that has been previously unaddressed.

At a broader level, we believe that our work highlights the need for a re-envisioning of the current state of affect detection through keystroke dynamics. The current research paradigm is to throw an assortment of classifiers in an attempt to identify particular affects in typing data. Our work, which has been conducted with a substantially higher degree of scientific rigor and experimental control as compared to prior work, was unable to identify any universal markers for stress nor was it able to identify groups of subjects with strongly shared markers. This finding strongly suggests that performing generic affect and stress detection in an across-subject manner – where one attempts to ascertain the affect present in a subject's typing data without having access to previous typing data from that subject – is unlikely to ever be successful. We suggest that researchers should more closely focus on within-subject classification, where previous typing data from the subject is available and classifiers can be personalized to individual subject; our work has demonstrated that classification accuracies above 90% is relatively straightforward in within-subject classification. Researchers who remain interested in across-subject classification would be best served by searching for universal markers or groups of subjects with strongly shared markers; this might be done by examining non-standard feature sets in keystroke dynamics or by incorporating other sources of information (e.g., mouse movements).

10.3   Comparison with existing work

We compare our work with both the existing literature on stress detection through keystroke dynamics and, more generally, on affect detection through keystroke dynamics. Much of the existing literature is insufficiently detailed about the experimental and analytical details of the work that it can be difficult to precisely ascertain what was done. However, to the best of our reading, the existing literature is largely focused on performing across-subject analysis (discussed in detail in Chapter 8), an understaking that we found wanting.

Attempting to perform across-subject analysis makes the implicit assumption that there are either universal markers or at least groups of subjects with strongly shared sets of markers. Without either of these, one can generally assume that the resulting analysis will be fruitless. Within the data collected in this thesis work, we found no evidence of either universal markers or groups of subjects with strongly shared markers. As would be expected, our across-subject analyses were largely poor; all classifiers we ran performed below 60% accuracy. When we compare our results with work from others focused on stress detection, reported classification accuracies are generally higher than what we obtained in our work, such as the 75% obtained by Vizer et al. (2009). Papers focusing on keystroke features themselves – such as work by Gunawardhane et al. (2013) and Kolakowska (2016) – generally report significant differences in keystroke features between neutral and stressed conditions; but, it is worth noting that these papers generally do not use vetted stressors and/or have independent validations of stress, so these results must be taken with only low confidence.

When we broaden our comparison to include papers focused on generic affect (not specifically stress) detection, the story remains largely the same. The reported classification accuracies – such as the 95.6% reported by Lv et al. (2008) – are higher than those that we obtained. Due to the lack of reported details, it is difficult to immediately discern why our results differ so significantly from those reported in the literature. We consider several possibilities.

Lack of experimental control. Much of the existing literature was performed under ill-specified and inexact experimental controls. Subjects were often exposed to unvetted stimuli with no assurance that subjects actually entered the desired affective states. Even when attempts were made to check the affective state of subjects, this was most often done via only self-reporting using instruments that are not previously vetted (unlike, for example, our STAI), without concomitant objective measurements of affect, such as blood pressure, heart-rate variability, and psychological inventories employed in our study. With such loose experimental controls, it is difficult for a reader to feel assured that the obtained classification results could be directly attributed to the supposed affect and not some other confounding variable.

Low sample size. More than half of the existing research into affect or stress detection via keystrokes has been conducted with sample sizes below 30. It is entirely possible that many positive results may simply be artifacts of small sample size. It is perhaps noteworthy that the largest study – with 100 subjects by (Tsihrintzis et al., 2008) – obtained results between 57% and 74%, a range that brackets our best results.

Inattention to potential confounding variables. A common confounding variable in keystroke research, and one that we paid a great deal of attention to in our own work, is the influence of practice. In a typical keystroke experiment, including the vast majority of those in the literature, subjects are asked to type an unfamiliar phrase into an unfamiliar piece of software using an unfamiliar keyboard and computer. Given these circumstances, it would not be surprising for subjects to exhibit a significant practice effect. Indeed, in our own work, we saw that many subjects had a substantial practice effect. The difference between our work and the work in the literature is that we incorporated a familiarization period where subjects could become largely practiced prior to providing the "real" data we would analyze; we also accommodated the practice effect by subtracting it out of our data. Without such a familiarization period, the first typing session would

generate highly unpracticed data while subsequent sessions would generate increasingly practiced data; these data would also be influenced by any affect present during their production. Even if affect induction was completely unsuccessful, with subjects remaining in a neutral state throughout the experiment, we would expect the generated data to be easily classified as unpracticed data is quite different from practiced data. This issue is magnified if only a small number of typing repetitions are collected as the repetition-to-repetition changes are highest when subjects just begin typing. We speculate that many of the high classification accuracies reported are merely the result of differentiation between unpracticed and practiced data and not actually differentiation between presence or absence of a given affect.

"File drawer problem". The term "file drawer problem" colloquially refers to the problem of bias being inserted into scientific literature by the tendency to publish only positive and/or confirmatory results while neglecting negative or contradictory results (which are then left to rot in the file drawer). From our own personal experience, keystroke dynamics is an area of research that seems to suffer from this problem. Much of the literature is filled with highly positive results, most of which cannot be independently replicated and for which data are not available. It is possible that the existing literature on affect/stress detection through keystroke dynamics suffers from the same problem; authors or editors chose to publish only the positive results, effectively hiding all of the negative ones.

10.4   Limitations and future work

As with any piece of research, our work is not intended to be the final word on stress detection through keystroke dynamics. Rather, this work was intended from the outset to be a proof-of-concept to determine whether stress manifests at all in typing data. We identify now a few limitations on our work and suggest natural directions in which the work could be extended.

Single stimulus. In our work, we use only a single stimulus string: `great friends are good to have`. While this string was carefully selected out of a pool of 100 candidate strings, it is nevertheless still only a single string. While we expect that the results obtained in this thesis, using this string, are likely to be representative of results using any string, this remains unknown until this work is reproduced using other stimulus items. It would be particularly interesting to allow subjects to type arbitrary text (i.e., a free-text experiment), as that would more closely resemble the natural environment where we would wish to perform stress detection through keystroke dynamics.

Single sitting. In our current study, each subject is run through the experimental protocol once. In the course of doing so, the subject provides a single typing episode in each of the baseline, stress, and recovery conditions. Since we have only a single episode for each subject, we have no method for ascertaining the consistency of his/her stress response over an extended period of time. Ideally, we would hope that this stress response is consistent across repeated applications of a stressor, but we cannot ascertain this without actually conducting such a study. We opted against performing such a repeated-measures study, as we expected that it would be too difficult for us to conduct. In addition to a proportionate increase in experimenter manpower, laboratory space, and financial compensation, we would expect that subject recruitment would be significantly more

difficult, especially if subjects had to refrain from relatively common behaviors (e.g., caffeine or alcohol consumption) for an extended period of time. There would also be the additional difficulty of finding appropriate relaxation and stressor stimuli with similar intensities; one might reasonably expect that repeated exposure to the same stimuli would have a smaller effect on the subject.

Generalizability. The results obtained in this thesis work focus on a limited subject population. Subjects were skilled and healthy typists, largely from a university campus. While we expect that these results would be similar for other subject populations, we cannot be certain without replicating the experiment with a different subject population.

Work must be reproduced/replicated. We hope the reader, at this point, has been convinced that significant attention was paid to the particulars of conducting this experiment and analyzing the resulting data. Nonetheless, some unknown flaw or quirk could have arisen in the course of our experiment that may significantly impinge on the validity of our results. Our intention is to make all of the data from this experiment available as soon as possible. We would hope that other researchers would see fit to reproduce and replicate this work to confirm our findings. We draw particular attention to this need since it is highly uncommon in keystroke dynamics to reproduce or replicate the work of others, despite this being common practice in other scientific disciplines.

Larger-scale experiments. We have drawn our conclusions in this study based on a subject pool of 116 subjects. One of the most notable findings was that subjects seem to be highly individualized in the markers for stress in their typing. An interesting question to posit is whether all subjects truly possess a unique set of markers for stress or if we merely have conducted an insufficiently large-scale experiment to see groups of subjects with similar markers. Such a question can only be answered with a large-scale version of experiment, perhaps with an order of magnitude more subjects.

Motion capture and/or pressure-sensitive keyboards. One of the limitations in keystroke dynamics research is that we are largely limited to analysis of hold and latency times and their derivative measures. Using motion capture devices or pressure-sensitive keyboards could potentially increase the information collected in a study, allowing researchers to analyze the 3-dimensional physical motions of typing rather than just the resultant hold and latency times. We wonder whether this richer information could allow for a better understanding of typing behavior, perhaps eventually resulting in higher classification accuracies in keystroke dynamics research.

Detecting other phenomena of interest. In this work, we have demonstrated that stress detection is possible through keystroke dynamics. Stress is not the only phenomenon of interest that one might wish to detect through typing behavior. With minor modifications, the experimental design, protocol, and analyses in this thesis could be applied to other affects or toward early detection and disease-progression tracking of afflictions such as Carpal Tunnel Syndrome, Alzheimer's disease, Parkinson's disease, dementia, or cognitive decline. We hope that the work performed in this thesis will enable and inspire others to investigate the non-computer-security applications of keystroke dynamics.

Despite the aforementioned limitations, as a proof of concept, our experiment and analyses seem sufficiently sound to confidently say that stress does manifest as changes to typing rhythms.

Chapter 11

Conclusion

At the outset of this work, our objective was to evaluate the promise of a new technology: detecting a user's stress through keystroke dynamics. This work was a proof-of-concept; we wanted to know if this technology was effective in an ideal setting. Throughout the course of designing and executing our study, we paid zealous attention to the particulars of the experiment so as to ensure the maximal validity of the study. It is our belief that our study is more rigorously-designed and scientifically-sound than any existing study on detecting stress or affect through keystroke dynamics.

Our work has shown that there is significant promise for stress detection through keystroke dynamics within a closed-world environment, where a computer system is aware of the identities of all persons on the system. Such systems are common in areas such as air-traffic control, nuclear power plant operation, government facilities, process-control operations in a typical facility, and many corporate and office environments. In our work, we demonstrated that reliable stress detection for a given user is possible so long as a system can be personalized to that user. Gathering the required data to perform this personalization is straightforward in a closed-world environment, where users are likely to use the same system for weeks, months, or even years. Our proof-of-concept work has shown that using simple off-the-shelf machine learning algorithms, it is possible to achieve nearly 90% classification accuracy rates. We imagine that such accuracies would be sufficiently high to be useful as one indicator of stress, perhaps alongside other independent indicators. With further refinements of the techniques used in this thesis work, we imagine that significantly higher accuracies would be possible; it may be possible to raise these accuracies high enough such that keystroke dynamics could function as the sole indicator of stress in a closed-world environment.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.

AD Instruments (2014). http://www.adinstruments.com/. Accessed: 14 January, 2014.

Agnan Kessy, Alex Lewin, K. S. (2016). Optimal whitening and decorrelation. *CoRR*, abs/1512.00809v4.

Alhothali, A. (2011). Modeling user affect using interaction events. Master's thesis, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada.

Amazon's Mechanical Turk (2014). https://www.mturk.com/mturk/welcome. Accessed: 10 November, 2014.

Andren, J. and Funk, P. (2005). A case based approach using behavioral biometrics to determine a user's stress level. In *Proceedings of the Workshop on CBR in the Health Sciences, The Sixth International Conference on Case-Based Reasoning (ICCBR-05)*, pages 9–17, Berlin, Germany. Springer. Presented at ICCBR-05 in Chicago, Illinois, USA.

Baker, N. A. and Redfern, M. S. (2005). Developing an observational instrument to evaluate personal keyboarding style. *Applied Ergonomics*, 36(3):345–354.

Bando, S., Nozawa, A., and Matsuya, Y. (2015). Multidimensional directed coherence analysis of keystroke dynamics and physiological responses. In *2015 International Conference on Noise and Fluctuations (ICNF)*, pages 1–4.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory (COLT 92)*, pages 144–152, New York, New York. ACM.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bryan, W. L. and Harter, N. (1897). Studies in the physiology and psychology of the telegraphic language. *Psychological Review*, 4(1):27–53.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, 24:386–396.

Critchlow, D. and Fligner, M. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation in glim. *Psychometrika*, 56(3):517–533.

Dickerson, S. S. and Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130(3):355–391.

Epp, C., Lippold, M., and Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. In *29th Annual Conference on Human Factors in Computing Systems (CHI 2011)*, pages 715–724, New York, NY. ACM. Presented at CHI 2011 in Vancouver, BC, Canada.

Fairhurst, M., Li, C., and Costa-abreu, M. D. (2015). Exploring emotion prediction from biometric-based keystroke dynamics data using multiagent systems. In *6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15)*, pages 1–6.

Foundation, P. S. (2018). Python software. https://www.python.org/. Accessed: 7 March, 2018.

Gaines, R. S., Lisowski, W., Press, S. J., and Shapiro, N. (1980). Authentication by keystroke timing: Some preliminary results. Technical Report RAND-R-2526-NSF, Rand Corporation, Santa Monica, CA.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Gunawardhane, S. D., Silva, P. M. D., Kulathunga, D. S., and Arunatileka, S. M. (2013). Non invasive human stress detection using key stroke dynamics and pattern variations. In *Proceedings of the 2013 International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 240–247. IEEE. Presented at ICTer 2013 in Colombo, Sri Lanka.

Gunetti, D. and Picardi, C. (2012). Keystroke analysis as a tool for intrusion detection. In Ahmed, A. A. E. and Taror, I., editors, *Continuous Authentication Using Biometrics: Data, Models, Metrics*, pages 193–211. Information Science Reference, Hershey, PA.

Hannan, D. (1999). Coral Sea Dreaming, Special Edition. David Hannan Productions, Mill Reef Entertainment. Motion picture on DVD. United States: DVD International. ISBN: 1-932198-7-6-8.

Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Hancock, P. and Meshkati, N., editors, *Human mental workload*, pages 139–183. Amsterdam, North Holland.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning*. Springer, New York, NY, USA, 2nd edition.

Hoffer, E. and Ailon, N. (2014). Deep metric learning using triplet network. *CoRR*, abs/1412.6622.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Khanna, P. and Sasikumar, M. (2010). Recognising emotions from keyboard stroke pattern. *International Journal of Computer Applications*, 11(9):1–5. Published By Foundation of Computer Science.

Killourhy, K. S. and Maxion, R. A. (2009). Comparing anomaly-detection algorithms for keystroke dynamics. In *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN-09)*, pages 125–134, Los Alamitos, California. IEEE Computer Society Press. Estoril, Lisbon, Portugal.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kolakowska, A. (2013). A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th International Conference on Human System Interactions (HSI)*, pages 548–555.

Kolakowska, A. (2015). Recognizing emotions on the basis of keystroke dynamics. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 291–297.

Kolakowska, A. (2016). Towards detecting programmers' stress on the basis of keystroke dynamics. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1621–1626.

Lakie, M. (2010). The influence of muscle tremor on shooting performance. *Experimental physiology*, 95(3):441–450.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Lehrer, P. M. (1987). A review of the approaches to the management of tension and stage fright in music performance. *Journal of Research in music Education*, 35(3):143–153.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lim, Y. M., Ayesh, A., and Stacey, M. (2014a). Detecting cognitive stress from keyboard and mouse dynamics during mental arithmetic. In *2014 Science and Information Conference*, pages 146–152.

Lim, Y. M., Ayesh, A., and Stacey, M. (2014b). Detecting emotional stress during typing task with time pressure. In *2014 Science and Information Conference*, pages 329–338.

Lim, Y. M., Ayesh, A., and Stacey, M. (2014c). The effects of typing demand on emotional stress, mouse and keystroke behaviours. In Arai, K., Kapoor, S., and Bhatia, R., editors, *Intelligent Systems in Science and Information 2014*, pages 209–225. Springer International Publishing.

Lim, Y. M., Ayesh, A., and Stacey, M. (2014d). Using mouse and keyboard dynamics to detect cognitive stress during mental arithmetic. In Arai, K., Kapoor, S., and Bhatia, R., editors, *Intelligent Systems in Science and Information 2014*, pages 335–350. Springer International Publishing.

Lim, Y. M., Ayesh, A., and Stacey, M. (2016). Exploring direct learning instruction and external stimuli effects on learner's states and mouse/keystroke behaviours. In *2016 4th International Conference on User Science and Engineering (i-USEr)*, pages 161–166.

Lovallo, W. R. (2005). *Stress & Health: Biological and Psychological Interactions*. Sage Publications, Thousand Oaks, CA, second edition.

Lv, H.-R., Lin, Z.-L., Yin, W.-J., and Dong, J. (2008). Emotion recognition based on pressure sensor keyboards. In *IEEE International Conference on Multimedia and Expo (ICME 2008)*, pages 1089–1092, Piscataway, NJ. IEEE. Presented at ICME 2008 in Hannover, Germany.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2017). *cluster: Cluster Analysis Basics and Extensions*.

Manuals Online (2018). Lifesource uc-322 digital scale. http://personalcare.manualsonline.com/manuals/mfg/lifesource/uc322.html. Accessed: 31 January, 2018.

Marteau, T. M. and Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, 31:301–306.

Maxwell, S. E. and Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Psychology Press, New York, New York, USA, 2nd edition.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81–97.

Miller, R. (1981). Simultaneous statistical inference. *SPRINGER-VERLAG INC., 175 FIFTH AVE., NEW YORK, NY, 1981, 300.*

Nahin, A. N. H., Alam, J. M., Mahmud, H., and Hasan, K. (2014). Identifying emotion by keystroke dynamics and text pattern analysis. *Behaviour and Information Technology*, 33(9):987–996.

Obaidat, M. S. (1995). A verification methodology for computer systems users. In *ACM Symposium on Applied Computing (SAC)*, pages 258–262, New York, NY, USA. ACM Press.

Open Broadcasting Software (2018). https://obsproject.com/. Accessed: 7 January, 2018.

Pachelbel, J. (1680). Pachelbel's canon. PWC 37, T. 337, PC 358.

Peacock, A. E., Ke, X., and Wilkerson, M. (2004). Typing patterns: A key to user identification. *IEEE Security and Privacy*, 2(5):40–47.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Purple Research Solutions (Accessed: 13 September, 2014). `http://www.purple-research.co.uk/framework.html`.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Respiration Belt (2014). https://www.adinstruments.com/products/respiratory-belt-transducer. Accessed: 14 January, 2014.

Ritter, F. E. and Schooler, L. J. (2001). The learning curve. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social & Behavioral Sciences*, volume 13, pages 8602–8605. Elsevier, Amsterdam, New York, 1st edition.

Samura, T. and Nishimura, H. (2009). Keystroke timing analysis for individual identification in Japanese free text typing. In *ICCAS-SICE 2009*, pages 3166–3170, Tokyo, Japan. SICE.

Saul, L. K. and Roweis, S. T. (2000). An introduction to locally linear embedding. Technical report, New York University.

Shikder, R., Rahaman, S., Afroze, F., and Islam, A. B. M. A. A. (2017). Keystroke/mouse usage based emotion detection and user identification. In *2017 International Conference on Networking, Systems and Security (NSysS)*, pages 96–104.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Spielberger, C. D., Gorsuch, R., Luchene, R., Vagg, P., and Jacobs, G. (1983). *Manual for the State-Trait Anxiety Inventory STAI (Form Y)*. Palo Alto, CA.

StarTech.com (2018). 4 port dual bus pci express usb 3.0 card with usap - lp4 + sata power. https://www.startech.com/Cards-Adapters/USB-3.0/Cards/PCI-Express-USB-3-Card-4-Dedicated-Channels-4-Port PEXUSB3S44V. Accessed: 31 January, 2018.

Stern, R. M., Ray, W. J., and Quigley, K. S. (2001). *Psychophysiological Recording*. Oxford University Press, New York, NY, 2nd edition.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643–662.

Teh, P. S., Teoh, A. B. J., and Yue, S. (August 2013). A survey of keystroke dynamics biometrics. *The Scientific World Journal*, 2013.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34(4):273.

Tsihrintzis, G. A., Virvou, M., Alepis, E., and Stathopoulou, I. (2008). Towards improving visual-facial emotion recognition through use of complementary keyboard-stroke pattern information. In *Proceedings of the 5th International Conference on Information Technology: New Generations*, pages 32–37, Los Alamitos, CA. IEEE Computer Society. Presented at Information Technology: New Generations in Las Vegas, NV, USA.

Vizer, L. M., Zhou, L., and Sears, A. (2009). Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies*, 67(10):870–886.

Wang, K., Wang, B., and Peng, L. (2009). CVAP: Validation for Cluster Analyses. *Data Science Journal*, 8:88–93.

Wetherell, M. A. and Carter, K. (2014). The multitasking framework: The effects of increasing workload on acute psychobiological stress reactivity. *Stress & Health*, 30(2):103–109.

Wickelmaier, F. and Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, and Computers*, 36(1):29–40.

Yu, E. and Cho, S. (2003). Novelty detection approach for keystroke dynamics identity verification. In Liu, J., Cheung, Y., and Yin, H., editors, *4th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2003)*, volume 2690 of *Lecture Notes in Computer Science (LNCS)*, pages 1016–1023, Berlin. Springer Verlag.

Zimmermann, P., Guttormsen, S., Danuser, B., and Gomez, P. (2003). Affective computing – a rationale for measuring mood with mouse and keyboard. *International Journal of Occupational Safety and Ergonomics (JOSE)*, 9(4):539–551.

Appendix A

A.1   Keystroke collection process

As typing data are generated, two separate logging processes occur; these generate a pair of output data files – MTPXML and Gizmo. When a user presses (or releases) a key, this information is first captured by the Gizmo hardware and then passed on to the operating system. The MTPXML data contains all keystrokes captured by the MTP program and uses the (relatively inaccurate) timestamps generated by the Windows XP operating system. Additionally, the MTPXML data also contains indications of when a user successfully completes a repetition of the phrase, makes an error, or is presented with a new screen (containing either instructions or a text box for a repetition of the phrase). The Gizmo data contains all keystrokes generated, regardless of which program these were sent to, and uses the highly accurate timestamps generated by the Gizmo hardware.

As its name implies, the MTPXML data are in XML format and are composed of "events". The majority of events are associated with the user pressing or releasing a key. However, the completion of a repetition, a typographical error, and the presentation of a new screen all generate events as well. Correctly typed repetitions – which are the only ones we are interested in – are bookended by a New Screen event one on end and a Correct Entry event on the other. As discussed in the next section, these events are crucial to converting the raw data files into a usable format. Figure A.1 depicts an example of the data format.

The Gizmo data are in a much simpler format, as depicted in Figure A.2 (right). Lines alternate between triplets containing a timestamp, an indication of a key-press (KEYDOWN) or key-release (KEYUP) event, and the ASCII code for the key pressed and lines containing duration information between these events. Note that the key events contained in the Gizmo file are a (generally strict) superset of the key events recorded in the MTPXML format; this is because the Gizmo data file contains ALL keys struck on the keyboard during a typing session, while the MTPXML contains only keys directed to the MTP software by the operating system.

A total of 4 pairs of MTPXML and Gizmo files are generated during the experiment – one each for the warmup, first neutral, stressed, and second neutral typing sessions.

A.2   Keystroke merger process

There are a total of three steps that must be performed to make the typing data usable from an analysis standpoint. First, any unwanted data or undesirable artifacts must be removed in the Keystroke Edit phase. The precise details will be discussed further in the next section, but such this phase involves removing data from rejected subjects and removing artifacts that impede the next two pre-processing steps. The second step is the Keystroke Merge phase, where a pair of

```
<event type="custom">
    <timestamp source="qpc">7.8026</timestamp>
    <name>NewScreenIndex</name>
    <value>2</value>
</event>
<event type="custom">
    <timestamp source="qpc">7.8099</timestamp>
    <name>ScreenDisplayed</name>
    <value>great friends are good to have</value>
</event>
<event type="keystroke">
    <timestamp source="ticks">7.8750</timestamp>
    <timestamp source="qpc">7.8692</timestamp>
    <virtual_key>VK_Return</virtual_key>
    <key>Return</key>
    <key_event>break</key_event>
</event>
<event type="keystroke">
    <timestamp source="ticks">8.7350</timestamp>
    <timestamp source="qpc">8.7305</timestamp>
    <virtual_key>VK_g</virtual_key>
    <key>g</key>
    <key_event>make</key_event>
</event>
```

Figure A.1: **MTPXML format.** A snippet of the raw MTPXML file format.

MTPXML and Gizmo files are merged together into a single file containing only the keystrokes seen by the MTP program but using the more accurate timestamps in the Gizmo file. The third step is the Phrase Table Generation phase, where the data in the merged format are converted into a standardized form suitable for keystroke analysis. We discuss each of these steps in turn.

Keystroke Edit. Keystroke editing can happen in both the MTPXML and Gizmo files, though it is much more common in the MTPXML file. The most common type of edit is the deletion of a entire session of collected data. Most deleted sessions are generated during the pre-experiment phase, when the experimenter starts MTP to ensure that it is working properly. By design, MTP creates a pair of MTPXML and Gizmo files whenever it is started; obviously, these files do not contain any useful information and are thus removed. Some deleted sessions are caused by typographical errors by the experimenter. As previously noted, MTP requires a subject number to be entered when it is started; if an error is made in typing the subject number, the experimenter will close MTP and restart it with the correct subject number. However, this will result in a pair of MTPXML and Gizmo files that must be discarded. Finally, sessions from rejected subjects – where the experiment was prematurely terminated due to subject inattention or because s/he did not meet the requirements of the study – are also removed in this step. Whenever a session is to be removed, for any reason, both files in the pair are removed.

The other type of edit that can be performed is to remove individual keystrokes from the data file. This is a step that is rarely taken, and is only used to remove keystrokes that are not of interest, but which are impeding the remainder of the pre-processing procedures. The most common

```
527.807816 KEYUP 82
3 255 0 0.1290
527.936803 KEYUP 69
1 191 0 0.0159
527.952727 KEYDOWN 65
3 251 0 0.1323
528.085075 KEYDOWN 84
1 255 0 0.0257
528.110756 KEYUP 65
3 255 0 0.1102
```

Figure A.2: **Gizmo format.** A snippet of the raw Gizmo file format.

keystrokes that are removed are extraneous keys (usually Backspaces) that occur during at the beginning of a repetition. As previously mentioned, correctly-typed repetitions are supposed to be preceded by a New Screen event and followed by a Correct Entry event. In between, there is the expectation that only keystrokes corresponding to the phrase "great friends are good to have" are present. However, due to a design defect in the original MTP program, keystroke events that occur immediately prior to the New Screen event can sometimes be logged after this event, breaking this expectation [1]. Such keystrokes must be removed so that the remainder of the pre-processing step will function. Finally, keystrokes may sometimes have to be removed because they occurred while the MTP program did not have focus. This is a highly rare occurrence, as the data are collected in a highly controlled environment; moreover, subjects are asked to place the mouse out of the way before they begin typing, reducing the possibility of accidentally clicking the mouse.

To make an edit, entries must be created in either the MTP-sedfile and/or the Gizmo-sedfile, depending on the file-type to be modified. These two files are in a custom format that serve as a "paper trail", providing a summary of all the edits that were made to the data; this is intended to provide a record of how the data were modified since their collection. Entries removing entire sessions of data require the specification of the subject number and the file-creation date and time for the file to be removed, though the latter can be omitted if all files from a subject are to be removed (i.e., if a subject is rejected for any reason). Entries involving deletion of a keystroke require specification of the subject number, file-creation date and time, and the precise timestamp of the keystroke of interest.

Keystroke Merge. The keystroke merger program takes in a pair of MTPXML and Gizmo files and outputs a merged file where all the MTPXML keystrokes now have the timestamps from the Gizmo file. The heart of the keystroke merger program is a longest-common-substring computation. With the exception of keystrokes right at the beginning and end of the file, which are usually associated starting or closing MTP, the merger program looks for a near-perfect match between the keys captured in the MTPXML file and those captured in the Gizmo file. The matched keystrokes are then written to an output file with the timestamps in the Gizmo file. Additionally,

---

[1]Technically speaking, the defect is due to two simultaneous threads logging to the same file. Meta-events, including New Screen and Correct Entry events, are generated by the Visual Basic MTP thread. Logging of the actual keystrokes is done by a separate thread before these are passed to the Visual Basic thread. Thus, a race condition exists where events in the log may not correspond to the order they were perceived by the Visual Basic thread, which evaluates the correctness of a typed repetition. Ultimately, removing this race condition would have required a redesign of the program; the condition manifests so rarely (approx. 5% of files) that it was easier to deal with in pre-processing.

the program also looks for correctly-typed repetitions of the phrase. Correctly-typed repetitions must be preceded by a New Screen event and ended by a CorrectEntry event, with precisely the expected keystrokes in between. In addition to outputting a merged output file, the keystroke merger program also outputs the number of correctly-typed repetitions it found. If these do not align with expectations [2], the file is manually examined and fixed by creating entries in the MTP-sedfile or Gizmo-sedfile. If such a misalignment occurs, the pre-processing procedure is fully restarted. Misalignments happen in approximately 3-5% of subjects.

Phrase Table generation. Once merged files have been generated for every single session of typing data, the phrase-table generation program is run. This program takes all of the merged files as input and outputs a single phrase table. Each row in the table corresponds to a single correctly-typed repetition; each column contains the duration of a single hold or latency time. This phrase-table generation program is relatively straightforward; it simply computes the hold and latency times for each of the correctly-typed repetitions identified in the Keystroke Merger phase. A small amount of meta-data (e.g., subject and session numbers) is also added by the program to help with bookkeeping.

A.3   Stimulus selection in keystroke dynamics

A.3.1   Problem and approach

The general problem is how to choose a suitable stimulus for an experiment in keystroke dynamics, based on the goals of the experiment. More specifically, in the present experiment that serves as the illustrative example case, how does one choose a stimulus string that is well-attuned to the keystroke task of detecting affect, based on the characteristics of typing rhythms? Our stepwise approach is four-fold: first, determine the criteria for the string; second, generate candidate strings; third, prune the list of candidates to a number aptly suited for analysis; and finally, select the best of the candidate strings. These steps will be detailed in the next section, and illustrated by a specific example in the sections on experimental methods.

A.3.2   Stepwise approach

Here we describe the stepwise method; implementation details are provided in the Method sections. **Candidate definition.** Defining what kind of stimulus is best suited to the research question simply involves asking what the keystroke experiment is intended to do. For example, such experiments can seek to distinguish among users (the most typical task), or to determine some other characteristic of the users, such as handedness or gender. Take handedness as an example. One could imagine that a simple string might serve best, where "simple" means no special characters, no upper case, etc. If most of the characters in the stimulus string are predominantly on one side of the keyboard, there won't be enough comparisons between keys struck on the left vs. the right to make an adequate comparison of how these respective keys are struck, and whether there are significant differences among them. So, a stimulus string that is attuned to the research task would be roughly evenly divided across the keyboard so as to attract keystrikes from both hands. There may be other constraints, as well, but this is just illustrative. While this stimulus might be effective for ascertaining handedness, it might not be the most effective stimulus for, say, gender. Every research question is associated with a keystroke task, and that task should use a stimulus string that is

---

[2]40 correctly-typed repetitions in the Warmup session and 80 repetitions in each of the three other session

best suited to answer the research question. This is not to say that a generic one-size-fits-all string would be unsuitable; just that a customized string might be better. Every aspect of the research question and keystroke task should be considered before settling on a definition, or requirement, for the stimulus string. Based on the requirements of the keystroke task, the stimulus string should be crafted to correspond to it, and a justification for the choice of stimulus should be provided.

Candidate generation. Given a definition for the stimulus string, it is likely that there are many suitable strings, although some might be better than others. One should generate a large number of candidate strings, and then choose the best from among them (called pruning, in the next step). The constraints provided in the candidate definition procedure will guide the generation process.

Candidate pruning. Once a set of candidate strings has been generated, each string should be analyzed individually for suitability as a final stimulus string. If the candidate set is too large for one-by-one scrutiny, then the set of candidates should be pruned in a principled way to extract a smaller, but higher quality, subset.

Candidate ordering/scaling. When the candidate list is suitably small, the candidate strings can be ordered from worst to best in accordance with the criteria in the definition. In our work, this ordering is done using a Thurstone scaling model, which determines a ranking of the phrases based on subjects' pairwise preferences among them.

A.3.3 Method – Candidate definition

In this section we go from the general to the specific, in terms of a concrete example in which a stimulus string needs to be "easy to type."

Our research question concerns detecting a user's affective state (stress) by examining changes to or characteristics of typing rhythm. One could imagine that these affective states might cause only subtle changes in typing – in the timings of key-holds and interkey latencies. Hence, we would want our users to type a stimulus string that induced as little noise, or variation, as possible, lest the noise mask the signal for which we are searching. As previously indicated, pilot studies in our lab have shown that easily-typed strings tend to have low variation. To help in defining the stimulus string, we consider the criteria shown in Table A.1.

Regarding the guidance in Table A.1, we have defined the requirements for our stimulus string as follows:

- English language, because our subject pool comprises native speakers of English.

- Memorable, so that users don't have to look at the screen more than once to apprehend the string (looking back induces pauses).

- Pre-habituated/familiar, so that rhythm doesn't change with repeated typings.

- All lower case; no punctuation or special characters. Mixed case, punctuation and special characters can be awkward to type, hence inducing pauses or breaks in rhythm.

- Time to type should be roughly 5 seconds so that enough repetitions of the string can be typed before the laboratory-induced affective state fully attenuates.

- Emotional content should be flat; no emotionally charged text. If the stimulus string itself induces emotion, it will be difficult to separate that from the effect of laboratory-induced affect.

- String length should be 30 characters; this is long enough to detect the sought-after affect, but short enough to remember, based on a glance.

- Minimal noise; variability in rhythm should be as low as possible from one string tying instance to another (for multiple repetitions of the same string).

- No constraints on keying patterns, which should not matter for this task.

- String type is words, because they are easier to remember that other kinds of strings; memorability is critical in this task.

- Content should comprise dictionary words, exclusive of proper nouns.

These constraints on the definition of our stimulus string leave us with the higher-level definition of 30-character, lower-case strings, comprised of English words, easy to remember and easy to type, the ease of typing imposed because easy-to-type strings will be intrinsically less variable, and hence less noisy. One example string would be: smell the sweetness of the rose.

A.3.4   Method – Candidate pruning

The objective of candidate pruning is to reduce a large pool of stimulus candidates to a smaller, more manageable pool. Because we planned to use pairwise comparisons for scaling the candidates, we pruned our candidate list to 20 phrases, for reasons (mainly due to resource limitations) explained in the next section.

We based the pruning on memorability - how memorable was each phrase, and which 20 of the 100 phrases were the most memorable? We used a memory task to ascertain which of the 100 candidate phrases were most easily remembered. In this task a subject viewed a phrase for 5 seconds, after which the phrase disappeared, and the subject was asked to type the phrase. The 20 phrases that were most often typed correctly were selected; they appear in Table A.2 in no particular order.

A.3.5   Method – Candidate ordering/scaling

This section describes how the 20 candidate phrases from the pruning procedure were ordered from easiest to hardest to type, and how the relative positions of the phrases were pinned to a scale. To achieve this, Thurstone scaling (Critchlow and Fligner, 1991; Thurstone, 1927) was applied to the 4129 pairwise comparisons amongst 20 phrases, provided by 413 Mechanical Turk subjects in an on-line experiment, whose procedure is given below.

The Thurstone model presumes that each phrase has an inherent "strength" that represents how easy it is to type; the easier a phrase is to type, the higher its strength. Given n different phrases ($n = 20$ in our case), we have n different strengths: $\mu_1, \mu_2, ..., \mu_n$. When a subject sees two phrases, he is more likely to express a preference for the stronger phrase. Moreover, the bigger the difference between the respective strengths of the two phrases, the more probable it is that the subject expresses a preference for the stronger of the two. In practice, what we observe are the counts of preferences from the subjects – how many times each phrase is preferred to each other phrase. Similarly, we observe the fraction of the time that each phrase is preferred to each other phrase; these fractions are denoted p_{ij} for $i, j \in \{1, ..., n\}$. The objective in fitting a Thurstone model is to find values for $\mu_1, ..., \mu_n$ that fit these fractions well.

What makes one fit better than another? In the way that linear regression makes an ideal assumption that the data lie on a line, $y = \beta \mathbf{x}$, with some noise getting in the way, the Thurstone model makes an ideal assumption that the relation $\phi^{-1}(p_{ij}) = \mu_i - \mu_j$ holds, with some noise

getting in the way; ϕ^{-1} here refers to the inverse of the normal CDF. In linear regression, one fit is better than another if it results in a smaller sum-squared-error:

$$\sum_{i=1}^{n}(y_i - \beta\mathbf{x}_i)^2$$

For a Thurstone model, one fit is also considered to be better than another if it results in a smaller sum-squared-error:[3]

$$\sum_{\substack{i,j \\ i\neq j}}(\phi^{-1}(p_{ij}) - (\mu_i - \mu_j))^2$$

The intuition behind fitting a Thurstone model is to assign strengths such that phrases with preference fractions close to 50-50 have similar strengths, while phrases with preference fractions that are skewed have considerably different strengths; the more skewed the fractions are towards a unanimous vote, the larger the difference there should be between the strengths of the two phrases.

Procedure. We conducted a Mechanical Turk (MTurk) (Amazon's Mechanical Turk, 2014) experiment with 413 subjects. The task presented 10 pairs of webpages to the subject. The first webpage in each pair contained two distinct phrases (of the 20) that subjects had to type correctly into text boxes. Once the two phrases were typed correctly, subjects could proceed to the next page where they were asked to indicate which phrase they found easier to type. Subjects saw each of the 20 phrases exactly once. Pairs of phrases were chosen in a pseudo-random fashion.

Outcome. The result of this process was a set of 4129 pairwise preferences, each of which indicated a subject's preference for one phrase being easier to type than the other – for example, the number of times that Phrase-1 was thought to be easier to type than Phrase-2, and so on, for all 20 phrases. These 4129 comparisons were the inputs to the Thurstone scaling procedure. The Thurstone scaling procedure was fitted using the eba (Wickelmaier and Schmid, 2004) package in R (R Core Team, 2012).

A.3.6 Results

Results are reported in terms of the ordering of the 20 phrases, from easiest to hardest to type, as well as the relative scaling amongst the phrases; and an empirical validation showing the difference between easy and hard phrases graphically.

A Thurstone model was fitted to the 4129 pairwise comparisons provided by 413 MTurk subjects, each of whom made 10 pairwise comparisons (except one, who provided only 9). A depiction of the fitted model is presented in Figure A.4. By convention, a Thurstone model gives the most preferred (easiest to type) phrase a value of 0, with all other phrases receiving more negative values. As can be seen in Figure A.4, subjects found phrase #1 (`great friends are good to have`) the easiest to type.

Ordering and scaling of phrases. Thurstone scaling produced an ordering and scaling of the 20 phrases. The ordering is shown in Table A.2, where the order is noted in the superscript at the

---

[3]In this work, we actually use the maximum-likelihood formulation of a Thurstone model, instead of the least-squares formulation we describe here. The difference between the two models is that the least-squares formulation breaks down (i.e., produces infinities) when there is a unanimous vote for one phrase over another, while the maximum-likelihood formation does not. We choose here to present the least-squares formation due to its much clearer intuition.
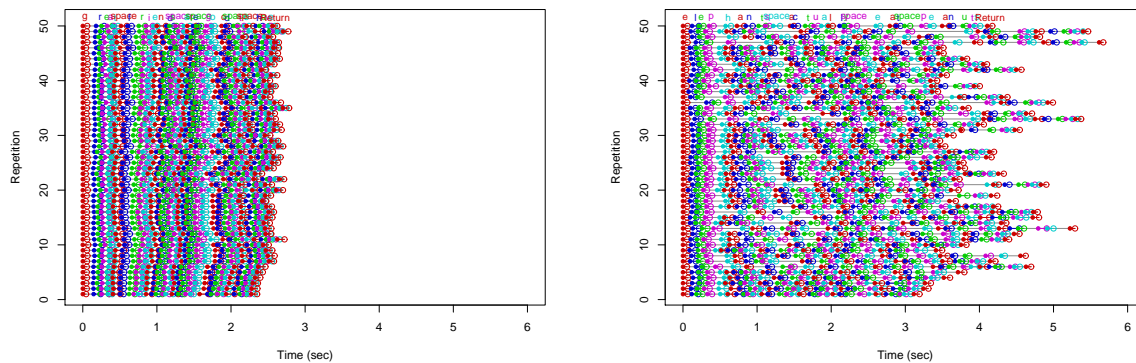
Figure A.3: **Easy vs. hard phrase.** Comparison between typing an easy phrase (left panel, great friends are good to have) and a more difficult phrase (right panel, elephants actually eat peanuts). Notice the low variability for the easy phrase vs. the high variability for the harder one. Both examples were typed by the same person, a skilled touch typist. Solid dots indicate the moment of key-press; open dots indicate key-release. Distance between like-colored solid/open dots is key-hold time. Open space is time between key presses.

end of each phrase. The easiest phrase to type was number 1, great friends are good to have. The hardest phrase to type was number 7, elephants actually eat peanuts. The scaling of the 20 phrases can be seen in Figure A.4. The figure shows the ordering of the phrases, as well as the relative distance among them.

Validation. Figure A.3 shows the typing rhythms of one subject in a laboratory validation of our result. The subject typed both the easy phrase (left panel) and the hard phrase (right panel). It is easy to see that the easy phrase engendered much less variation (noise) in terms of both total typing time (x-axis) and latency times.
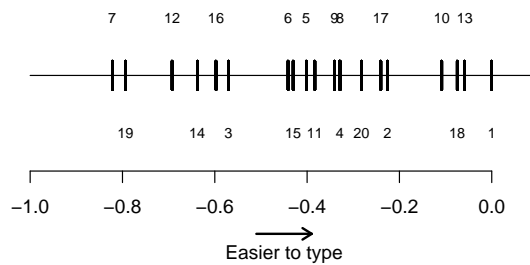
Figure A.4: **Thurstone ordering/scaling.** Depiction of the ease of typing a phrase, as determined by a Thurstone model fit to data from touch-typists. Both no-look and sometimes-look touch-typists had their data included in this model. Numbers 1-20 correspond to phrases, as listed in Section A.3.4. By convention, the easiest phrase (#1 – `great friends are good to have`) is given a value of 0, with all other phrases given more negative values. As examples, the second easiest phrase is #13, with the hardest phrase being #7.

Language: English, French, Italian, etc.

Memorability: the string should be "memorable" in that once the string is apprehended, a subject should be able to type it without looking back to be reminded of the next character or word. Recall that human short-term memory is limited to 7 ± 2 "chunks", whether they are characters, words, phrases or anything else that can be encoded in memory within about seven elements Miller (1956).

Habituation: the more often you type a string, the better you get at it, and hold/latency timings will change. To avoid mistaking these changes in time for indications of affect, we prefer a string that is resistant to habituation; this suggests that the string should be pre-selected as one that is already well habituated, such as "the" in English. Of course three characters is too short, but the string should be as resistant to habituation as possible.

Case: mixed case is harder to type and more variable; all upper case is harder to read, and hence might affect readability of the given string.

Punctuation: periods, commas and apostrophes can be awkard to type (more so than not typing them), so we opt for no punctuation.

Special characters: characters such as @ or # or $ can be hard to type, especially when mixed within running text.

Typing time: The amount of time it takes to type the stimulus.

Emotional content: string should not induce emotion in the typist; e.g., no distressing content.

Length of string: must be long enough to facilitate the detection of affect, yet not violate constraints such as memorability; a longer string will more likely show the effect of affect, but will also be less memorable.

Noise: by noise we mean high variability; the string must induce as little variability as possible.

Required keying: At least 5 transitions between the left and right hand for a proper touch typist.

String type: letters, phrases, sentences, free text, fixed text, PIN.

Word content: Only English words found in the dictionary; no proper nouns.

Strikingly unfamiliar words: Syzygy or uglify, for example, by their very unfamiliarity, can induce timing changes that may be mistaken for signs of affect.

Character content: All letters in the alphabet, selected letters only; inclusion and exclusion criteria.

Table A.1: Stimulus selection considerations.

1. great friends are good to have[1]
2. where is the smallest donation[13]
3. jeans are not very comfortable[18]
4. he is going to an art festival[10]
5. he hates seeing spiders inside[2]
6. my cat did not enjoy the water[17]
7. elephants actually eat peanuts[20]
8. she goes to the football games[8]
9. he played games in high school[4]
10. it was too nice to stay inside[9]
11. she studied very hard at night[11]
12. always say please to be polite[5]
13. there is no need to argue here[15]
14. spilling milk is very bad luck[6]
15. the bride and groom are lovely[3]
16. their first apartment was tiny[16]
17. please arrive on time tomorrow[14]
18. they could see fire in the sky[12]
19. celebrate your accomplishments[19]
20. spring has more hope than fall[7]

Table A.2: **Experimental phrases.** These 20 phrases, used in the Mechanical Turk experiment, were given to the scaling algorithm in the order shown. The algorithm produced the order shown in superscripts. For example, "great friends" was the easiest to type, "where is" was 13th and "elephants actually" was hardest (20th).

A.4   Recruitment materials

The vast majority of the subjects in our study were recruited when they contacted us after viewing recruitment posters around campus. The posters prompted the potential subject to send mail to our experimenter. Our experimenter would then send an initial e-mail to the potential subject to ascertain whether s/he met the inclusion and exclusion criteria. Once the potential subject responded, a stock acceptance or rejection mail was sent. These recruitment materials are enclosed below in the following order: 1) recruitment poster, 2) initial e-mail, 3) acceptance e-mail, and 4) rejection e-mail.

Seeking Volunteers for a Research Study

Principal Investigator: Dr. Roy Maxion

The purpose of this study is to establish whether cognitive load can be detected through the way somebody types. It is well-known that a person under heavy cognitive load tends to make more errors. This is particularly problematic for individuals working in highly-sensitive environments, such as air-traffic controllers or nuclear power plant operators. If we can detect cognitive load through changes in typing, then we may be able to prevent catastrophic errors from happening.

To participate in this research you must:

- **Be at least 18 years of age**
- **Be a fluent English speaker**
- **Have at least 3 years of typing experience on a computer**
- **Be able to type at least 30 words per minute**

You must NOT:

- **Have a history of cardiac, neurological, anxiety, stress, or sleep disorders**
- **Have hypertension – blood pressure above 140/90**
- **Have had a stroke**
- **Suffer from color-blindness**

Participation in this study involves:

- **A single 3-3.5 hour experimental session on the CMU campus**
- **You will receive up to $60 as compensation for your participation**

To find out more about this research study, please contact:

- Patricia Loring (Dr. Maxion's research assistant) at sawako@cs.cmu.edu (Please include "Research Study" in the subject line)

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Subject: Research Study
Email: Patricia Loring
sawako@cs.cmu.edu

Hi, <name>.

Thank you for your interest in our study of the effects of
cognitive loading on typing rhythms. Before we schedule an
appointment for your joining the study, we'd like to confirm your
eligibility to participate. Below are a few statements; please
read these statements carefully, and REPLY to this email, answering
each question with an 'x' in the appropriate "true/false/not-sure" box.

Please note that we will be verifying these statements again on
the day of the study. If you are ineligible, we'll need to
excuse you from the study, without compensation.

 1. I am at least 18 years old.
      True [ ] False [ ] Not sure [ ]

 2. I speak English fluently.
      True [ ] False [ ] Not sure [ ]

 3. I have at least three years of experience typing on a computer.
      True [ ] False [ ] Not sure [ ]

 4. I can type at least 30 words per minute. (Typing at 30 words
    per minute means you can type the sample text below in 1 minute.)
      True [ ] False [ ] Not sure [ ]

 5. I have no history of cardiac disorders.
      True [ ] False [ ] Not sure [ ]

 6. I don't have any history of neurological disorders.
      True [ ] False [ ] Not sure [ ]

 7. I do not have any history of anxiety or stress disorders.
      True [ ] False [ ] Not sure [ ]

 8. I've never had a stroke.
      True [ ] False [ ] Not sure [ ]

9. I'm not currently being treated by a doctor for a sleep disorder.
   True [ ] False [ ] Not sure [ ]

10. I don't suffer from any form of color-blindness.
    True [ ] False [ ] Not sure [ ]

11. My blood pressure is below 140/90. (If you have a blood
    pressure above 140/90, you may have hypertension.)
    True [ ] False [ ] Not sure [ ]

12. I have not heard anything about this experiment beyond
    what is on the recruitment materials.
    True [ ] False [ ] Not sure [ ]

Once we hear from you, and have confirmed your eligibility for
this study, we will contact you again for your appointment.

Sincerely,

  Patricia Loring

Sample text - typing the text below, in one minute, is 30 words/min:

  The old man is wearing a ship captain's uniform with a red cloth
  in his back pocket. He is hunched forward as he gazes out at the
  beautiful blue ocean.

<Acceptance letter for subjects - 021416@1815>
<Subject field of this email should indicate topic and subject name.>
<Don't forget to attach the map.>

Hi, <name>.

We are happy to let you know that you are eligible to participate in our study on typing rhythms and cognitive loading. We are currently able to schedule you for the following dates and times:

     <list of dates and times (both beginning and end times)>

Can you please let me know which date and time would be most convenient for you, by REPLYing to this email?

The study session will held in Gates Hall, room 8122, at Carnegie Mellon University. Directions to campus and to the specific building and room are given below, at the end of this email.

Please note that you ...

1. Must not consume psychoactive drugs (e.g., anti-depressants, marijuana, ecstasy, LSD, Ritalin, etc.) during the 48 hours before the start of the study.

2. Must not consume alcoholic beverages for 48 hours before the start of the study.

3. Must not consume excessive caffeine (more than 3 cups of coffee, or equivalent, in a day) for 24 hours before the start of the study.

4. Must not consume ANY caffeine or other stimulants for 2 hours before the study.

5. Must wear a loose top, such as a half-sleeve/short-sleeve t-shirt on the day of the study. If it is more convenient for you, you are welcome to bring a loose top, and change into it once you arrive. (The loose top facilitates easy attachment of electrocardiogram leads and blood pressure cuff.)

You will receive 10 dollars compensation for participating in the

study. You will have the opportunity to earn an additional 50 dollars by being highly engaged during the experiment, bringing your total compensation as high as 60 dollars. Compensation will be provided in the form of Giant Eagle gift cards (which we figure everyone can use).

If you have any questions or concerns, you can reach me at (412)-268-5628 or by email at sawako@cs.cmu.edu. I look forward to seeing you soon!

Sincerely,

Patricia Loring

Directions to CMU

A campus map (in PDF format) has been attached to this e-mail.

 - Yellow circle: the Gates building at CMU.
 - Red circles: Pittsburgh PAT bus stops near the CMU campus.
 - Purple circle: CMU parking garage.

Getting to the CMU campus by bus:

The 61A, 61B, 61C, 61D, 58, 67, 69, and 28X buses all stop at Forbes and Morewood. These stops are marked in red on the enclosed campus map.

Getting to the CMU campus by car:

The CMU campus is located at 5000 Forbes Avenue. On-campus parking is available in the East Campus Garage (purple circle on the map); you can enter the garage from the intersection of Beeler and Forbes. Once you have parked, exit the garage out onto Forbes Avenue, in the direction of downtown. The Gates building is roughly one long block away on the same side of the street; it will be on your left (yellow circle on map).

Getting to Gates Hall:

The CMU campus can be difficult to navigate, so we recommend that when you arrive on campus, you consult the attached map, and/or ask the person nearest to you: How do I get to Gates Hall?

If you lose your way, or you are running late, please call Patricia Loring (412-268-5628) to get help or let her know.

Hi <name>,

Unfortunately, you do not meet the eligibility requirements for participation in this study. However, we would like to thank you for your interest in the study!

Sincerely,

Patricia Loring

A.5 Multi-tasking framework instructions

In addition to the verbal instructions provided by the experimenter, we also provide subjects with a written version of the same instructions, which are included below.

P		YELLOW
Retrieve List		
True / False		

2

Target with concentric circles, red dot center. Reset / Reset

3	1	4	4
1	5	5	3
1	3	3	1
5	4	2	3
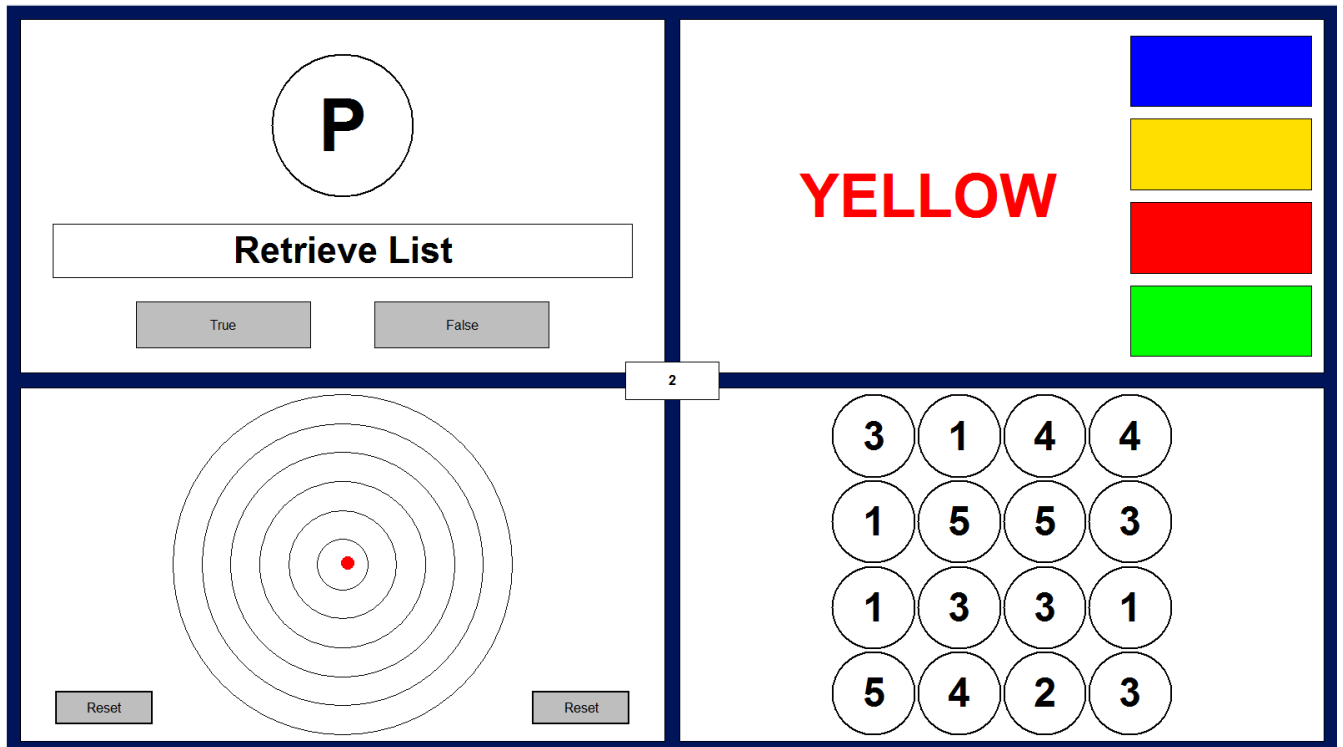
Instructions

1. This exercise consists of 4 different tasks, one in each quadrant of the screen.
2. Your goal is to be as fast and as accurate as you can be, on **ALL** of the tasks at the same time.
3. Your score is located in the center of the screen.
4. Doing well at a task will gain you points, while doing poorly will cause points to be deducted from your score.
5. Failing to attend to tasks in time will result in them timing out, causing you to lose many points.
6. Your performance will be evaluated based on 1) your score, 2) whether you attend to **ALL** of the tasks, and 3) whether you are going quickly enough.
7. The four tasks are described on the following pages.
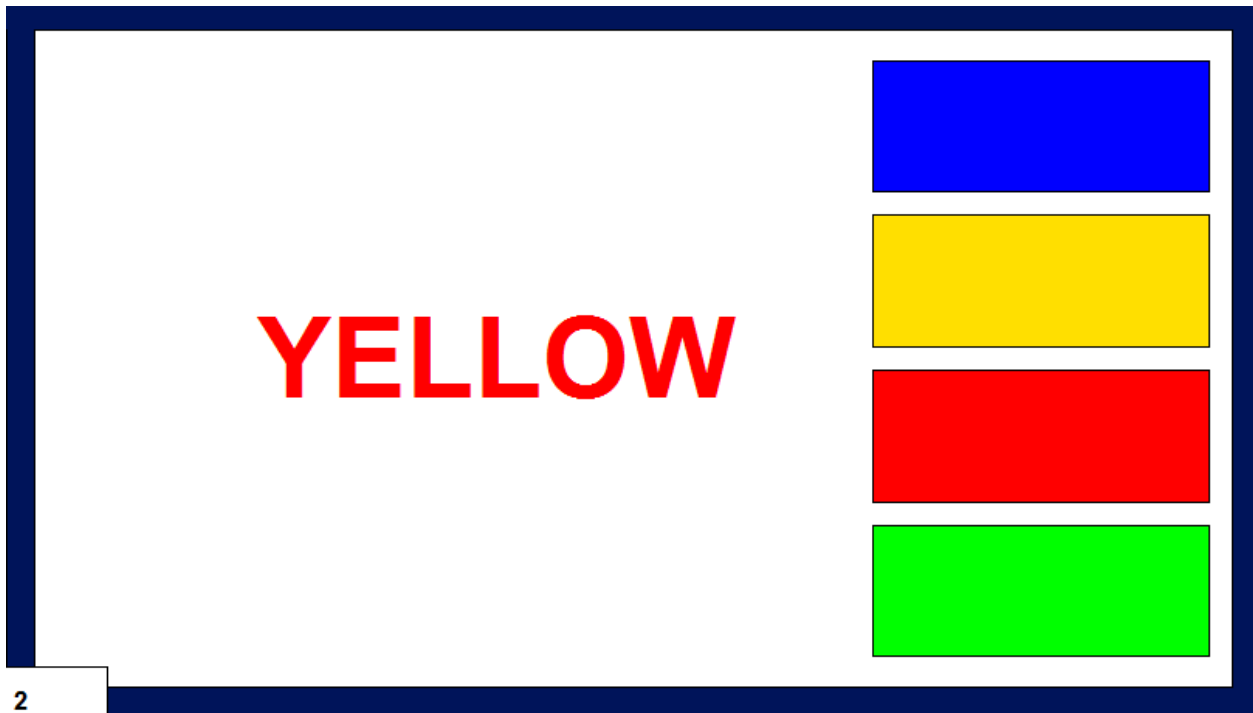
Color Word Task

1. Your goal is to click the colored box that corresponds to the **font color** of the displayed word.
2. In the above example, you should click on the red box, since the word "YELLOW" is written in a red font.
3. You will be awarded 10 points for a correct response. You will lose 10 points for an incorrect response.
4. **Failing to respond quickly enough will cause a timeout, causing you to lose 30 points.**
5. After each answer, or after you fail to respond quickly enough, a new word color will be displayed.
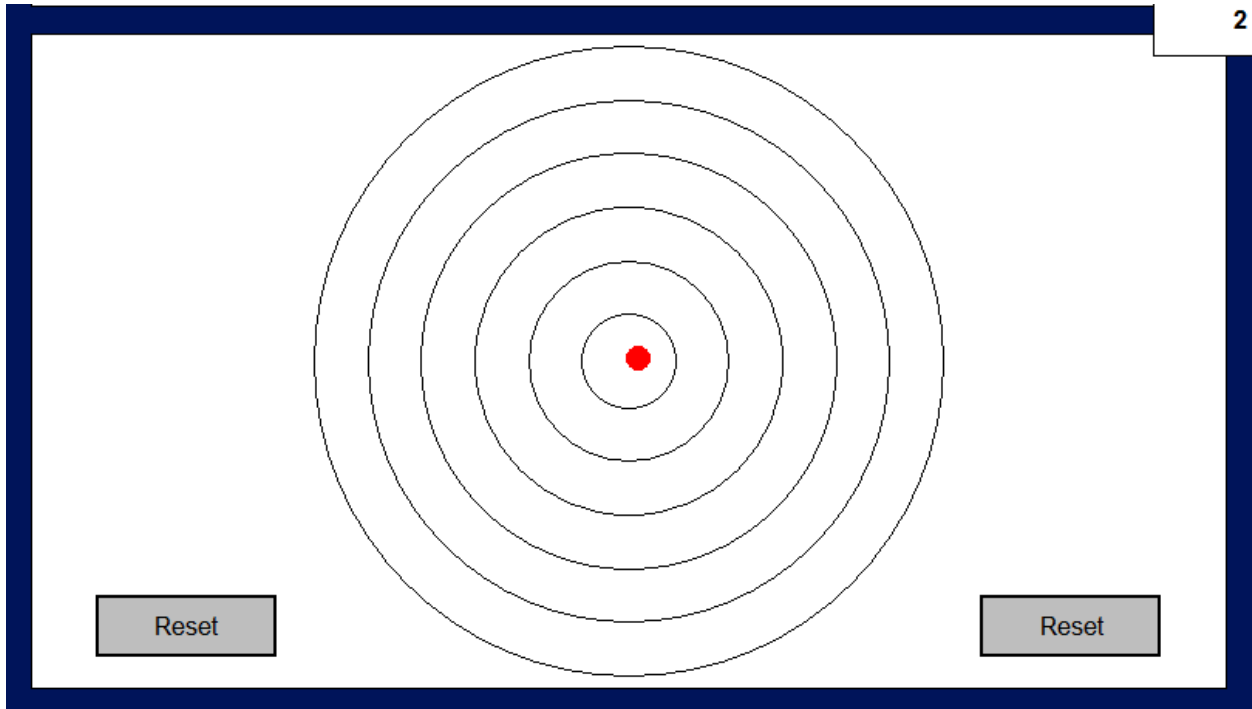
Target Tracker Task

1. Your goal is to keep the red dot inside the target.
2. The dot will start at the center of the target, and move outward.
3. Clicking the "Reset" button will re-position the dot to the center of the target.
4. The further out the dot is when you click Reset, the more points you will get (up to 10 for the outermost ring).
5. **If the dot drifts past the outer ring, the task will timeout, causing you will lose 10 points every half second.**
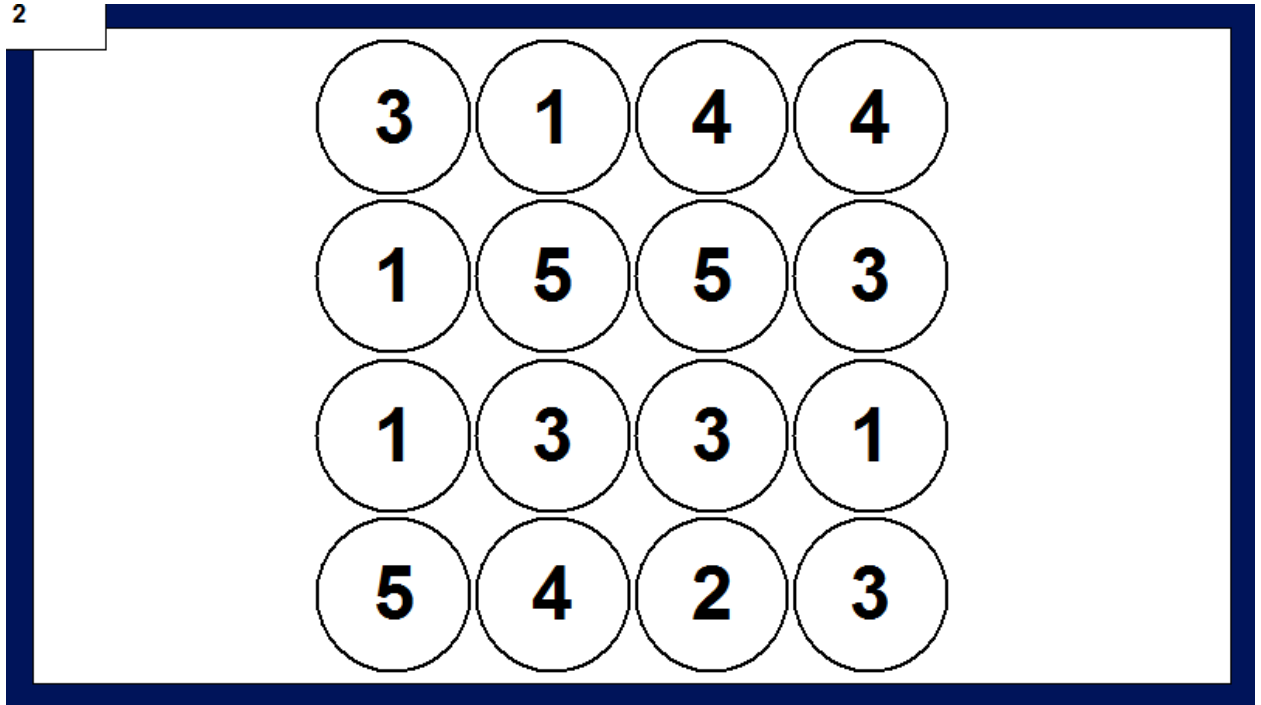
Highest-Number Task

1. Your goal in this task is to highlight all of the circles that contain the highest number (5 in this example).
2. Click a circle to highlight it. Click it again to un-highlight it.
3. In this example, the highest number is 5, so you should click all of the circles that contain 5.
4. You get 10 points once you've highlighted all the circles that contain the highest number.
5. **The task will timeout if you do not highlight the circles quickly enough, causing you to lose 30 points.**
6. The grid will reset when you finish highlighting all the circles, or after you have taken too long.
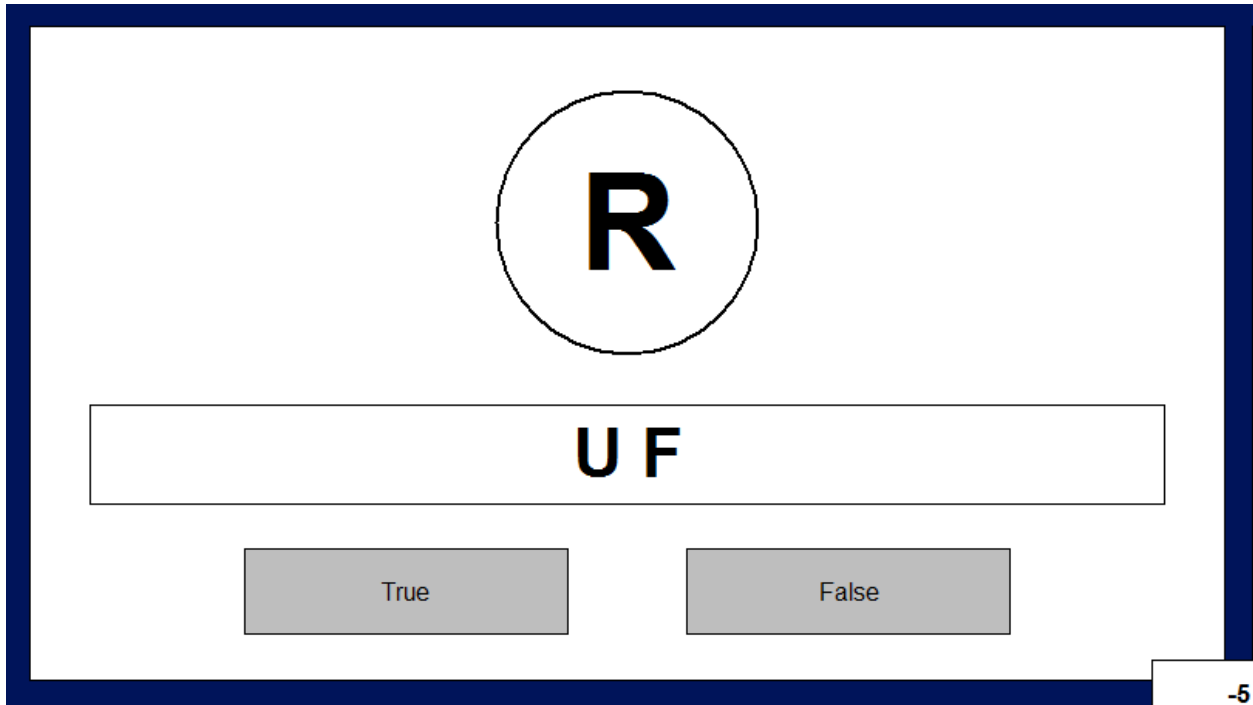
Letter-Match Task

1. Your goal in this task is to indicate whether the letter in the circle is contained in the list of letters in the horizontal box. The list can be different and longer than in the example.
2. The letters in the box will initially be visible for several seconds, after which they will be replaced by "Retrieve List".
3. If you cannot remember the letters in the box, you may click on "Retrieve List" to see them again.
4. You can only respond to the task when the box contains the words "Retrieve List."
5. Respond by clicking the True or False buttons, depending on whether the circled letter was in the box. In the above example, you should click False, because R is not contained in the horizontal box.
6. You gain 10 points for a correct answer, and lose 10 for an incorrect answer.
7. **The task will timeout if you do not respond quickly enough, losing you 30 points.**
8. Clicking "Retrieve List" will cost you 5 points.

A.6 Experiment forms

As referenced in Chapter 5, a number of forms are used during the course of our experiment. The forms used are the Long-form STAI (Y-2), PSS-10, Short-form STAI, and NASA-TLX. We also use a demographic survey that is commonly administered in our lab. For the sake of transparency, we reproduce each of these forms in the coming pages.

ID					STAI_Long_Y2
Initials					
DOB	m	m	d	d	yy
Date	m	m	d	d	yy
Session					

SELF-EVALUATION QUESTIONNAIRE : STAI Long Form Y-2 (Trait)

DIRECTIONS:

A number of statements which people have used to describe themselves are given below. Read each statement and then circle the appropriate number to the right of the statement to indicate how you generally feel. There are no right or wrong answers. Do not spend too much time on any one statement, but give the answer which seems to describe your present feelings best.

	ALMOST NEVER	SOMETIMES	OFTEN	ALMOST ALWAYS
21. I feel pleasant	1	2	3	4
22. I feel nervous and restless	1	2	3	4
23. I am satisfied with myself	1	2	3	4
24. I wish I could be as happy as other seem to be	1	2	3	4
25. I feel like a failure	1	2	3	4
26. I feel rested	1	2	3	4
27. I am "calm, cool, and collected"	1	2	3	4
28. I feel that difficulties are piling up so that I cannot overcome them	1	2	3	4
29. I worry too much over something that really doesn't matter	1	2	3	4
30. I am happy	1	2	3	4
31. I have disturbing thoughts	1	2	3	4
32. I lack self-confidence	1	2	3	4
33. I feel secure	1	2	3	4
34. I make decisions easily	1	2	3	4
35. I feel inadequate	1	2	3	4
36. I am content	1	2	3	4
37. Some unimportant thought runs through my mind and bothers me	1	2	3	4
38. I take disappointments so keenly that I can't put them out of my mind	1	2	3	4
39. I am a steady person	1	2	3	4
40. I get in a state of tension or turmoil as I think over my recent concerns and interests	1	2	3	4

The following questions ask you about your feelings and thoughts **during the last month**.
For each question you will be asked to indicate how often you felt or thought a certain way.
The best approach is to answer each question fairly quickly.

For each question circle one of the following answers:

0 = Never
1 = Almost Never
2 = Sometimes
3 = Fairly Often
4 = Very Often

In the last month……

		Never				**Very Often**
1	How often have you been upset because of something that happened unexpectedly?	0	1	2	3	4
2	How often have you felt that you were unable to control the important things in your life?	0	1	2	3	4
3	How often have you felt nervous or stressed	0	1	2	3	4
4*	How often have you felt confident about your ability to handle your personal problems?	0	1	2	3	4
5*	How often have you felt that things were going your way?	0	1	2	3	4
6	How often have you found that you could not cope with all the things that you had to do?	0	1	2	3	4
7*	How often have you been able to control irritations in your life?	0	1	2	3	4
8*	How often have you felt that you were on top of things?	0	1	2	3	4
9	How often have you been angered because of things that happened that were outside of your control?	0	1	2	3	4
10	How often have you felt difficulties were piling up so high that you could not overcome them?	0	1	2	3	4

ID			*STATE_ANX_VAS*				
Initials							
Date							
DOB							
Session							

A number of statements that people have used to describe themselves are given below.
Read each statement, and then mark on the line at the most appropriate point to indicate
how you feel right now, at this moment.

I feel calm

not at all   very much

I feel tense

not at all   very much

I am upset

not at all   very much

I feel relaxed

not at all   very much

I feel content

not at all   very much

I feel worried

not at all   very much

ID				*NASA-TLX*
Initials				
DOB				
Date				
Session				

Please tick-mark each line at the point which matches your
experience with the task you have just completed.

A) **MENTAL DEMAND** – How much mental and perceptual activity was required (e.g.,
thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or
demanding, simple or complex, exacting or forgiving?

    Low ————————————————————————— High

B) **PHYSICAL DEMAND** – How much physical activity was required (e.g., pushing, pulling,
turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack
or strenuous, restful or laborious?

    Low ————————————————————————— High

C) **TEMPORAL DEMAND** – How much time pressure did you feel due to the rate or the pace
at which the tasks or task elements occurred? Was the pace slow and leisurely, or rapid and
frantic?

    Low ————————————————————————— High

D) **EFFORT** – How hard did you have to work (mentally and/or physically) to achieve your
level of performance?

    Low ————————————————————————— High

E) **PERFORMANCE** – How successful do you think you were in accomplishing the goals of the
task set by the experimenter (or yourself)? How satisfied were you with your performance
in accomplishing these goals?

    Poor ————————————————————————— Good

F) **FRUSTRATION** – How insecure, discouraged, irritated, stressed and annoyed versus
secure, gratified, content, relaxed and complacent did you feel during the task?

    Low ————————————————————————— High

Keystroke Experiment Demographic Survey

The objective of this survey is to help us understand the kinds of things that influence a person's typing style. Your data will be kept confidential. Some items are marked "Reserve;" ignore them.

General

Gender: ☐ Male ☐ Female

Write your age here (or check one age group): _____

☐ 18-20 ☐ 21-25 ☐ 26-30 ☐ 31-35 ☐ 36-40

☐ 41-45 ☐ 46-50 ☐ 51-60 ☐ 61-70 ☐ 71-80

(Reserved) _____

Highest level of education that you have completed:

☐ High-school/GED ☐ Some college

☐ 2-year college degree ☐ 4-year college degree - Major:_____

☐ Some graduate school ☐ M.A. or M.S - Major:_____

☐ Ph.D. - Area of thesis:_____

Current occupation:

☐ Undergraduate student Major:_____

☐ Master's student Field:_____

☐ Ph.D. student Field:_____

☐ Staff ☐ Administrative / ☐ Technical

☐ Faculty Field: _____

☐ Other Specify: _____

A.7   Protocol

The experiment conducted as part of this thesis work was far more complicated than previous experiments that we had run in our lab. To execute the experiment properly and consistently, we developed a detailed protocol laying out the precise steps required in the experiment. The basic layout of the protocol is simple, there is an instruction to perform on the left, with checkboxes on the right that should be marked when the instruction is completed. For simpler tasks, there are sometimes multiple instructions corresponding to a single checkbox. At times, the checkbox is replaced by a circle indicating a button to be pressed on a piece of equipment (e.g., the print button on the blood pressure machine) or a conditional instruction (e.g., skip to the debriefing step if a subject is not eligible). Actions that should be performed by the experimenter are in regular font, verbal instructions are in *italics* and particularly noteworthy instructions (as deemed by the experimenter herself) are in **bold**.

We reproduce here the full protocol document. We do so both for the sake of transparency and in the hope that others might find it helpful in their own research.

Subject number: _____   Date: _____   Time: _____

1.0	**Equipment**	
1.1	Ensure the following experiment equipment are available:	
	1.1.1	Still camera on tripod	Complete? ☐
	1.1.2	Remote for the still camera	Complete? ☐
	1.1.3	Subject's display, keyboard, mouse, gizmo	Complete? ☐
	1.1.4	3 large lamps on clamps, remote for top lamp	Complete? ☐
	1.1.5	4 webcams	Complete? ☐
	1.1.6	Green camera background screen	Complete? ☐
	1.1.7	Case of Coke to serve as a footrest for shorter subjects	Complete? ☐
	1.1.8	Speakers for Experimenter's laptop	Complete? ☐
	1.1.9	Experimenter's laptop	Complete? ☐
	1.1.10	Flash drive for saving experiment data	Complete? ☐
	1.1.11	**Electrodes - 3**	Complete? ☐
	1.1.12	Electrode leads	Complete? ☐
	1.1.13	Respirator belt	Complete? ☐
	1.1.14	PowerLab	Complete? ☐
	1.1.15	Critikon blood pressure machine with 2 different cuff sizes	Complete? ☐
	1.1.16	Pulse transducer	Complete? ☐
	1.1.17	Timer	Complete? ☐
	1.1.18	Photomon, keyboard, screen, mouse	Complete? ☐
	1.1.19	Labmon, keyboard, screen, mouse	Complete? ☐

	1.1.20	UPS for Labmon and Photomon	Complete? ☐
	1.1.21	UPS for Experimenter's laptop	Complete? ☐
	1.1.22	Scale for measuring weight	Complete? ☐
	1.1.23	Samsung Digital Presenter	Complete? ☐
	1.1.24	Laptop connected to Digital Presenter	Complete? ☐
	1.1.25	Chair for subject	Complete? ☐
	1.1.26	Chair for experimenter	Complete? ☐
Notes			

2.0	**Signs/Images/Labels**			
2.1	Ensure the following signs/images/labels are available:			
	2.1.1	Experiment in progress sign - on the door		Complete? ☐
	2.1.2	Height scale - on door		Complete? ☐
	2.1.3	Photo of hand with blue dots - on wall		Complete? ☐
	2.1.4	**Participant Subject # label for hand photos** - **on Digital Presenter**		Complete? ☐
	2.1.5	Image of EKG electrode attachment - on wall		Complete? ☐
	2.1.6	Image of Respirator placement - on wall		Complete? ☐
	2.1.7	Image of EKG electrode leads attachment - on wall		Complete? ☐
	2.1.8	Screen shot of appropriate Labchart readings - on wall near Labmon		Complete? ☐
	2.1.9	**Participant Subject # label for desk** - on desk		Complete? ☐
	2.1.10	**Participant Subject # label for KPECS** - on desk		Complete? ☐
	2.1.11	A schedule on the desk with the sequence of experiment steps so the subjects know where they are in the experiment		Complete? ☐
Notes				

3.0	**Supplies in Box**		
3.1	Ensure the following supplies are available in the box:		
	3.1.1	Scotch tape for Subject # labels	Complete? ☐
	3.1.2	Paper for the subject to write on	Complete? ☐
	3.1.3	Ball point pens	Complete? ☐
	3.1.4	Hypoallergenic gloves	Complete? ☐
	3.1.5	Alcohol swabs for removing electrode residue on skin and cleaning equipment	Complete? ☐
	3.1.6	Masking tape to ensure wires are secured	Complete? ☐
	3.1.7	Blue dot stickers for hand photos	Complete? ☐
	3.1.8	Blue dot stickers for KPECS	Complete? ☐
	3.1.9	4 AA batteries for the scale	Complete? ☐
	3.1.10	4 AAA batteries for the still-camera remote	Complete? ☐
	3.1.11	Hair ties or hairpins	Complete? ☐
	3.1.12	Tissues	Complete? ☐
	3.1.13	Bottle of water	Complete? ☐
3.2	Separate box for subject's cell phone and pocket contents		Complete? ☐
Notes			

4.0		**Manila Folders**	
4.1		Ensure the following are in the manila folder labeled with Subject #:	
	4.1.1	Printout of protocol which has the subject number and date on each form	Complete? ☐
	4.1.2	Consent form, complete with 1. **Highlighting** where the **subject** needs to **initial and sign** 2. **Experimenter signature**	Complete? ☐
	4.1.3	Demographic survey with subject number and date	Complete? ☐
	4.1.4	1 copy of the Long-Form STAI Y-2 with subject number and date	Complete? ☐
	4.1.5	1 copy of the PSS-10 with subject number and date	Complete? ☐
	4.1.6	6 copies of the VAS form with subject number and date	Complete? ☐
	4.1.7	6 copies of the NASA-TLX with subject number and date	Complete? ☐
4.2		Ensure the following are in the Cog Load - Purple Instructions manila folder:	
	4.2.1	Copy of the **Purple instructions**	Complete? ☐
4.3		Ensure the following are in the Cog Load - Subject Payment Form manila folder:	
	4.3.1	List of participant IDs, names, and schedule	Complete? ☐
	4.3.2	**Subject payment record form**	Complete? ☐
	4.3.3	**Payment for subject**	Complete? ☐
Notes			

5.0	**Video Check**			
5.1	Ensure that the video is working properly on Photomon[1]:			
	5.1.1	Double click on the desktop shortcut: **Command Window**		Complete? ☐
	5.1.2	Ensure the command window path is: **C:\Users\Administrator\Desktop**		Complete? ☐
	5.1.3	Start the cameras in command window: **start-cameras.bat tdemo**		Complete? ☐
	5.1.4	Ensure that all cameras start [2].		Complete? ☐
	5.1.5	Right click mouse button on the bottom of the screen and select **Show windows stacked**.		Complete? ☐
	5.1.6	The experimenter will check the videos on Photomon are displaying: 1. **Right camera** - **keyboard** 2. **Top camera** - **keyboard**, **labels** 3. **Left camera** - **keyboard** 4. **Face camera** - **green background only**		Complete? ☐
	5.1.7	Stop all cameras with: **stop-cameras.bat**		Complete? ☐
Notes				

---

[1]Photomon: login: administrator password: adm~020297

[2]Do not install updates if you are prompted. Inform Shing-hon after the experiment is over that there are updates to be installed.

6.0	**LabChart Check**	
6.1	**Turn on the Powerlab** (button is in the back).	Complete? ☐
6.2	Ensure that LabChart is working properly on Labmon[3]:	
	6.2.1	On the Desktop, double click the shortcut: **Stress_expt**.	Complete? ☐
	6.2.2	Ensure miniwindow is visible: **Preset Comment Function Keys**	Complete? ☐
	6.2.3	Attach the **pulse transducer** to left thumb.	Complete? ☐
	6.2.4	Start recording in LabChart by pressing: **Start**	Complete? ☐
	6.2.5	Plug the pulse transducer into the **PowerLab Slot 1**.	Complete? ☐
	6.2.6	Check the screen to ensure that you have a clean reading.[4]	Complete? ☐
	6.2.7	Plug the pulse transducer into the **PowerLab Slot 2**.	Complete? ☐
	6.2.8	Ensure that you have a clean reading.	Complete? ☐
	6.2.9	If either **PowerLab Slot 1** or **PowerLab Slot 2** does not produce a clean signal, continue to try higher-numbered slots until two working slots are found. Attach the respiration belt input and the EKG input to the two working slots.	
	6.2.10	Stop recording by pressing the button: **Stop**.	Complete? ☐
	6.2.11	Quit LabChart by clicking on the upper right **X button**. Do not save the file.	Complete? ☐
	6.2.12	Ensure that all EKG and respiration belt wires are untangled.	Complete? ☐
6.3	**Turn off the Powerlab.**	Complete? ☐
Notes		

---

[3]Labmon: login: administrator password: adm~020295

[4]A clean reading will have hills for the heartbeats. You may have to adjust the range to see the signal clearly. This can be done by clicking on the **drop-down arrow** next to the channel name and selecting one of the options. A good range setting will allow you to clearly see the signal without the signal ever going off the charts.

7.0	**MTP check**	
7.1	Ensure that MTP is working properly:	Complete? ☐
	7.1.1	Turn on the experimenter's laptop.	
	7.1.2	Open a command window and type: **run-cog-warmup.bat tdemo**.	
	7.1.3	Close the MTP window by clicking on the upper right **X button**.	
	7.1.4	In the command window, type: **q**.	
Notes		

8.0	**Purple check**		
8.1	Ensure that Purple is working properly:		Complete? ☐
	8.1.1	Turn on the experimenter's laptop if it's not on.	
	8.1.2	Open a command window and type: **run-purple.bat**.	
	8.1.3	Purple should display the main screen.	
	8.1.4	Close the Purple window with **Ctrl + q**.	
Notes			

9.0	**Hand Photo Check**		
9.1	Ensure that the digital presenter and laptop is working properly:		Complete? ☐
	9.1.1	Plug in and turn on the DP.	
	9.1.2	Turn on the laptop attached to the DP.	
	9.1.3	Double-click the Digital Presenter shortcut on the laptop and click on: **Full**.	
	9.1.4	Create two files for subject: **s000-FirstL-left-month-day-year** and **s000-FirstL-right-month-day-year**.	
	9.1.5	Quit the digital presenter program and turn off the DP.	
Notes			

10.0	**Scale Check**	
10.1	Ensure that the scale is working properly:	Complete? ☐
	10.1.1	Power on the scale and step on it to ensure it takes a reading.	
	10.1.2	If the scale will not turn on, replace the batteries; spare batteries are in the box of supplies.	
Notes		

11.0	**Still Camera Check**	
11.1	Ensure that the still camera is working properly:	
	11.1.1	Ensure that the legs of the tripod are on the floor markings.	Complete? ☐
	11.1.2	Unplug the charger and plug back in the remote. **BE CAREFUL! The mini-USB is very delicate and prone to bending!**	Complete? ☐
	11.1.3	Power on the camera and the remote.	Complete? ☐
	11.1.4	Take a picture with the remote.	Complete? ☐
	11.1.5	Ensure that a picture is taken and the subject's chair is in focus.	Complete? ☐
	11.1.6	Turn off the still camera and the remote.	Complete? ☐
Notes		

12.0	**Room Set Up**	
12.1	Set up the room:	
	12.1.1	Turn on all 3 LED lights and ensure that the subject area is illuminated.	Complete? ☐
	12.1.2	Turn on the Powerlab (button is in the back).	Complete? ☐
Notes		

13.0	**Subject Arrival**	
13.1	On the door, the experimenter will put up the sign: **Experiment in Progress**.	Complete? ☐
13.2	The experimenter will **silence the ringer for the phone**.	Complete? ☐
13.3	Greet the subject:	Complete? ☐
13.3.1	*Hi, I'm Pat. Please come in.* Check to see if the subject is wearing **a skintight top or a heavy sweater**[5].Then, ask the subject to: 1. Put down their **book bag** 2. Hang up their **coat or jacket** 3. Turn off their **cell phone** 4. Get rid of their **chewing gum or a mint** because we don't want it to impact their typing.	
13.3.2	*Now please have a seat here and make yourself comfortable.* Then, 1. Explain how to **adjust seat height, armrests** 2. Offer the coke case **as a footrest** 3. Tell them to empty their pockets, jewelry into the **box**. 4. Tell them we want them to be **comfortable typing**.	
13.3.3	Introduce the equipment to the subject. Mention the keyboard, monitor, cameras, LED lights, and backdrop.	
Notes		

[5]If they are, ask if they have anything they can change into. If not, let them know they may not be able to participate in the experiment.

14.0	**Subject Briefing**	
14.1	Brief the subject:	Complete? ☐
	14.1.1	*The purpose of this experiment is to see whether cognitive load can be detected through the way somebody types.*	
	14.1.2	*This experiment will take approximately 3-3.5 hours.* 1. *Start with some forms* 2. *Chance to get use to the experiment equipment* 3. *Two rest periods where you'll watch videos* 4. *Assorted computer tasks* 5. *Compensation after all tasks are complete*	
	14.1.3	*Do you have any questions?*	
Notes			

15.0	**Consent Form**	
15.1	The experimenter will present the consent form to the subject and obtain his/her signature:	
	15.1.1	*The first step in this experiment is to obtain your **informed consent**.*	
	15.1.2	*As part of this, I'm going to first ask you a series of yes and no questions to confirm that you are eligible to participate.* [6]	
	15.1.3	Are you at least **18 years old**?	**Yes**
	15.1.4	Do you **speak English fluently?**	**Yes**
	15.1.5	Do you have at least **three years of experience typing** on a computer?	**Yes**
	15.1.6	Can you type at least **30 words a minute** [7]?	**Yes**
	15.1.7	Do you have any history of cardiac disorders?	**No**
	15.1.8	Do you have any history of neurological disorders?	**No**
	15.1.9	Do you have any history of stress or anxiety disorders?	**No**
	15.1.10	Have you ever had a stroke?	**No**
	15.1.11	Are you currently being treated by a doctor for a sleep disorder?	**No**
	15.1.12	Do you suffer from any form of color-blindness?	**No**
	15.1.13	Is your blood pressure **below 140/90** [8]?	**Yes**
	15.1.14	Have you consumed **any alcoholic beverages in the past 48 hours**?	**No**
	15.1.15	Have you consumed **more than 3 caffeinated beverages in the past 24 hours**? A caffeinated beverage could be a cup of coffee, a can of soda, or energy drinks.	**No**
	15.1.16	Have you consumed **any caffeine or other stimulants in the past 2 hours**?	**No**
	15.1.17	Have you consumed **any psychoactive drugs, such as anti-depressants, Ritalin, or drugs like marijuana or LSD, in the past 48 hours** [9]?	**No**
	15.1.18	Have you heard **anything about the experiment other than what is on the recruitment poster**?	**No**

---

[6]If the subject does not meet ALL of the criteria listed here, inform them that they are not eligible to participate in the study.

[7]If the subject is not sure, show them the sample text and ask if they think they could type that in a minute.

[8]Both the systolic and diastolic blood pressure must be below the limit.

[9]A psychoactive drug is any drug that might affect your mood, cognition, or perceptive ability, including antidepressants, marijuana, ecstasy, LSD, Ritalin, etc.

	15.1.19	**Finally, are you feeling well today?** *Is there anything that might affect your typing, such as a minor cold or a flare-up of carpal tunnel?* [10]	Complete? ☐
	15.1.20	**IF THE SUBJECT IS ELIGIBLE FOR THE STUDY**, then say *Great, looks like you are eligible for the study!*	Complete? ☐
	15.1.21	*Now, about the compensation:* 1. *We will provide compensation for your time at the completion of this study.* 2. *It will be up to 60 dollars* 3. *In the form of Giant Eagle gift cards* 4. *10 dollars for completing the study* 5. *50 dollar bonus for* **being focused and performing well**[11]	Complete? ☐
	15.1.22	*Your participation in this study is completely voluntary. You are free to leave at any time, but if you leave early, you will only* **receive 10 dollars**.	
	15.1.23	*Please take a moment to read through the form now. I need your* **initials on page 3**, *and* **your signature** *on the last page.*	
	15.1.24	Ensure that the subject: 1. **Initialed all three places on page 3** 2. **Signed and dated the form for the Participant Signature**	Complete? ☐
	15.1.25	Ensure the experimenter signed and dated the form for the **Signature of Person Obtaining Consent**.	Complete? ☐
	15.1.26	Place the consent form in the manila folder.	Complete? ☐
	15.1.27	**IF THE SUBJECT IS NOT ELIGIBLE FOR THE STUDY**, say *I'm afraid that you're not eligible for the study. We can still provide you 10 dollars of compensation for showing up today, but we can't continue with the experiment.* Have subject sign **Subject payment record form**. [Proceed to the debriefing section on page 53 for subject payment instructions.]	DEBRIEF
Notes			

---

[10] Write the response in the notes box below. If the subject is feeling fine, right that down.

[11] If a subject asks what this means, say that s/he must be attentive to all instructions, pay attention to the task at all times, and perform well on all tasks.

16.0	**Blood Pressure Reading**	
16.1	Take a blood pressure reading from the subject:	Complete? ☐
16.1.1	**Turn on the Critikon.**	
16.1.2	*Now, I am going to take your blood pressure.*	
16.1.3	Remove any clothing that will prevent taking an accurate blood pressure reading.	
16.1.4	Place the cuff on the upper arm. Make sure: 1. the **air is squeezed out of the cuff** 2. the **muscles are not tense** 3. the BP cuff has **some overlap of velcro** 4. the **BP cuff artery arrow is pointed at the pulse** (it's slightly off-center and about 1" from the elbow crease) 5. the **cuff index line falls within the range markings**. If it's too small/big, replace the cuff with the appropriate size.	
16.1.5	Then, check to ensure: 1. the **tightness of the cuff allows 1-2 fingers to slip in** 2. the **velcro attachments are secure** 3. the **hose connected to the cuff is not kinked or warped**	
16.1.6	Press **Inflate/Stop**.	
16.1.7	Press **Print** but **do not** tear off the printout on the machine–you'll do this at the end.	
16.1.8	Systolic: _____ Diastolic: _____ Pulse: _____ MAP: _____	**PRINT**
16.1.9	Acceptable BP: $\leq$ **140 systolic / $\leq$ 90 diastolic**. *It looks like you are in the acceptable range.*	
16.1.10	Unacceptable BP: **NOT $\leq$ 140 systolic and NOT $\leq$ 90 diastolic**. *It looks like your blood pressure is outside the acceptable range. Let's have you relax for 5 minutes before I take another reading to see if you fall inside the range.* After 5 minutes, take another reading:	
16.1.11	Systolic: _____ Diastolic: _____ Pulse: _____ MAP: _____	**PRINT**
16.1.12	If the second reading is acceptable, and continue with the rest of the experiment.	

	16.1.13	If it is not accepable, inform the the subject that their blood pressure does not meet the requirements for our study and they cannot be included. Give them the $10 gift card for showing up. Have subject sign **Subject payment record form**. [Proceed to the debriefing section on page 53 for subject payment instructions.]	
	16.1.14	**Remove the blood pressure cuff** from the subject.	Complete? ☐
	16.1.15	**Turn off the Critikon.**	Complete? ☐
Notes			

17.0	**Demographic Survey**	
17.1	The experimenter will ask the subject to fill out the **demographic survey**:	
	17.1.1	Give a copy of the demographic survey and a pen to the subject.	
	17.1.2	*Next we have a demographic survey.*	
	17.1.3	*Please fill out your* **initials and date of birth** *at the top of the first page.*	
	17.1.4	*You can just ignore questions that are marked:* **"Reserved"**.	
	17.1.5	**Check** the demographic survey to make sure the subject did not skip any questions.	Complete? ☐
	17.1.6	Place the demographic survey into the manila folder.	Complete? ☐
Notes			

18.0	**Long-form STAI**	
18.1	The experimenter will ask the subject to fill out the **Long-form STAI** to the subject:	
	18.1.1	Give a copy of the Long-form STAI to the subject.	
	18.1.2	*Here's the next form. Please circle the number corresponding to statement describing* **how you generally feel.**	
	18.1.3	*Don't worry about your initials and date of birth, I will copy that later.*	
	18.1.4	Allow the subject to respond to all questions on the form.	
	18.1.5	**Check** that the subject has responded to all questions on the form.	Complete? ☐
	18.1.6	Place the long-form STAI into the manila folder.	Complete? ☐
Notes		

19.0	**PSS-10**	
19.1	The experimenter will ask the subject to fill out the **PSS-10**:	
	19.1.1	Give a copy of the PSS-10 to the subject.	
	19.1.2	*And now here's the last form.*	
	19.1.3	*For each question, please circle a number to indicate* **how often you felt or thought a certain way in the past month**.	
	19.1.4	**Check** that the subject has responded to all questions on the form.	Complete? ☐
	19.1.5	Place the PSS-10 into the manila folder.	Complete? ☐
Notes		

20.0	**Hand Photos**	
20.1	Take two hand photos of the subject:	
	20.1.1	**Turn on the DP and the black laptop**.	
	20.1.2	On the black laptop, click the desktop icon: **Samsung Digital Presenter**.	
	20.1.3	Enlarge the View Panel by clicking: **Full**	
	20.1.4	*Now I'm going to take your hand photos.*	
	20.1.5	*Please come over here and sit in this chair.*	
	20.1.6	*First, I am going to* **stick these dots on your hands** *like this photo.* [Experimenter shows the blue stickers and gestures to photo.]	
	20.1.7	Apply the blue stickers to the subject's hands.	
	20.1.8	*Please place your hand so the tip of your middle finger is covering the blue square.*	
	20.1.9	Take the photo of the right hand.	
	20.1.10	Save the photo of the right hand into the Hand Measurement folder. Name the file: **s000-FirstL-right-month-day-year**	
	20.1.11	*Now, switch hands.*	
	20.1.12	Take the photo of the left hand.	
	20.1.13	Save the photo of the left hand into the Hand Measurement folder. Name the file: **s000-FirstL-left-month-day-year**	
	20.1.14	Double check that the images look ok.	Complete? ☐
	20.1.15	*Please* **remove the dots on your hands** *and toss them in the trash.*	Complete? ☐
	20.1.16	Turn off and unplug the DP; close the laptop.	
Notes		

21.0	**Weight and height measurement**		
21.1	Take weight and height measurements:		
	21.1.1	*Next, we're going to measure your weight so please* **remove your shoes**.	
	21.1.2	Turn on the scale.	
	21.1.3	*Please don't hold onto or lean against anything while you're on the scale.*	
	21.1.4	**Weight:** _____ [12]	Complete? ☐
	21.1.5	*Next we are going to measure your height so please stand with your heels against the door.*	
	21.1.6	**Height:** _____ [13]	Complete? ☐
	21.1.7	*Go ahead and put your shoes back on and take a seat.*	
Notes			

---

[12] Make sure that the subject is not leaning against or on anything. (kg)

[13] No shoes. Write down the number that appears on the tape. (in) Actual height will be this + 18 in.

22.0	**Subject Familiarization**	
22.1	The experimenter will say the following:	Complete? ☐
	22.1.1	*Next up, I'm going show you the forms and software that youll be using for the rest of the experiment, so that you can get comfortable with them.*	
Notes		

23.0	**NASA-TLX and VAS**		
23.1	The experimenter introduces the NASA-TLX and VAS:		Complete? ☐
	23.1.1	Present a copy of the **NASA-TLX** and a copy of the **VAS** to the subject.	
	23.1.2	*Here are two forms that I will ask you to fill out after various tasks.*	
	23.1.3	NASA-TLX: *On this form you need to make a vertical mark to indicate* **how you felt about the task you just completed.** *For each item, you should be responding to the bold item.*	
	23.1.4	VAS: *On this form you need to make a vertical mark to indicate* **how you feel at the moment**.	
	23.1.5	*Do you have any questions?*	
Notes			

24.0	**KPECS**	
24.1	Attach KPECS dots to the subject.	Complete? ☐
	24.1.1	*Next, I'll show you the software that we'll use to collect your typing data.*	
	24.1.2	*While you practice using the software, I'll be taking a few pictures with a still camera to help us capture your typing posture.*	
	24.1.3	*I need to attach a few stickers to you now so that we have some reference points.* [Show the subject the KPECS picture.]	
	24.1.4	Confirm that the **ear hole** is visible.	
	24.1.5	Attach a dot to the **shoulder.**	
	24.1.6	Attach a dot to the **elbow knob.**	
	24.1.7	Attach a dot to the **wrist knob.**	
	24.1.8	Attach a dot to the **hip.**	
	24.1.9	Power on the camera and the remote.	
	24.1.10	Look through the camera to see if the angle and focus is good.	
	24.1.11	Ensure that the remote is next to the experimenter laptop.	
Notes		

25.0	**MTP**		
25.1	The experimenter introduces MTP:		
	25.1.1	Type at the experimenter's laptop command prompt: **run-cog-warmup.bat** <**subject-number**>.	Complete? ☐
	25.1.2	**Move cursor** to the corner of the screen.	
	25.1.3	*This is the program that we will be using to collect your typing data.* Explain: 1. In this experiment, you'll type the same phrase repeatedly. 2. For this practice session, youll type the **same phrase 40 times**. 3. For the experiment itself, you'll **type the phrase 80 times**. 4. After you type the phrase, you must press the **Return key**. 5. This **counter** increments when you type the phrase correctly. 6. In this experiment we want to collect your **natural typing style**. 7. Not a **speed** nor an **accuracy** contest. So try to settle into a normal, comfortable pace. 8. If you make a typo, you will hear a '**ding**' and the text box will **gray out** so type the phrase from the beginning again. 9. Timing is critical in this experiment so please **don't take breaks** in the middle of typing a phrase. **If you need to stretch or talk**, please do so after you have typed in the entire phrase including the Return key.	Complete? ☐
	25.1.4	*One final thing -* **Please keep both your feet flat on the floor while you type.** *This is essential for the BP and EKG readings.*	Complete? ☐
	25.1.5	*Do you have any questions? Please start now.*	
	25.1.6	Take **3 KPECS photos** while the subject is typing.	Complete? ☐
	25.1.7	When the subject is done, close the MTP window and in the laptop command window, type: **q**	
	25.1.8	*Please go ahead and* **take off the dots** *and give them to me.*	Complete? ☐
	25.1.9	**Turn off the still camera and the remote.**	Complete? ☐
Notes			

26.0	**Purple**	
26.1	The experimenter introduces Purple:	
	26.1.1	*Next, we'll go over the Purple software which requires you to multi-task. You'll get to practice using it for* **2 minutes** *now. In the experiment, however, you'll be using it for* **15 minutes** *and the tasks will be a lot harder. Please pay attention because this is* **important**.	
	26.1.2	[Hand the subject the Purple instruction packet.] *Here are the instructions for the tasks. I will go over them with you.*	
	26.1.3	Provide Purple overview: 1. Perform 4 tasks simultaneously 2. Correct responses give you points 3. Incorrect responses lose you points 4. **Timeouts** will lose you **lots of points** 5. Must aim to be as fast and accurate on **ALL** of the tasks 6. I will **evaluate** your performance throughout 7. Your **score** will be displayed in the center of the screen	Complete? ☐
	26.1.4	Provide letter task overview: 1. This is a memory task 2. Random string of letters in box 3. Letters hidden by **"Retrieve List"** 4. When to respond true 5. When to respond false 6. Give an example 7. **Cannot respond while letters are visible** 8. Lose points to reveal letters by clicking on "Retrieve List" 9. Better to lose some points by revealing the letters than to get answers constantly wrong	Complete? ☐
	26.1.5	Provide Stroop overview: 1. Respond by clicking on the **font color** 2. Give an example	Complete? ☐

	26.1.6	Provide target overview:
		
		  1. Dot moves outwards quickly
		
		  2. Reset button to recenter dot
		
		  3. More points the further out the dot is
		
		  4. **Continuously lose points if it is outside furthest circle**
			Complete? ☐

	26.1.7	Provide number grid overview:
		
		  1. Must click on all copies of the **biggest number**
		
		  2. Give example
		
		  3. Can click again to unselect
		
		  4. Once you're done, the grid will refill itself
			Complete? ☐

	26.1.8	*Go ahead and read through the rest of the instructions now. Let me know if you have any questions.*	
26.2		While the subject is reading the instructions, load up Purple.	
	26.2.1	Type at the experimenter's laptop command prompt: **run-purple.bat**.	Complete? ☐
	26.2.2	Start Purple with the 2 minute demo configuration file: **Warmup.cfg**	Complete? ☐
	26.2.3	When the subject is done reading, ask *Are you ready to start?*	
	26.2.4	Have subject **move the Mouse** to a comfortable position.	Complete? ☐
	26.2.5	**Please keep both your feet flat on the floor** *while you use Purple. Please start now.*	Complete? ☐
	26.2.6	Observe the subject during the 2-minute period and ensure that they are performing each task correctly. If you see a subject repeatedly making mistakes at one task, offer corrective guidance.	
	26.2.7	Once the 2-minute period is over, give the subject the option of going for another 2-minute period. *Would you like to practice for a few more minutes?* Repeat this process until both you and the subject are satisfied that the subject fully understands the tasks.	
26.3		*Great, we're now done with the familiarization period.*	Complete? ☐
Notes			

27.0	**Attach Sensors**	
27.1	Inform the subject that we will be attaching sensors to him/her:	Complete? ☐
27.1.1	*In a moment, I will be attaching some sensors to you.* Explain: 1. These sensors will monitor your physiological signals. 2. Once the sensors are attached, they need to stay attached to you for the rest of the experiment. 3. So you must **stay in your chair for the next two hours or so**. 4. If we have to remove the sensors, then we will have to **abort the experiment and we will only be able to partially compensate you.** 5. So, would you like to take a break now to get a drink of water or use the restroom?	
Notes		

28.0	**Explanation of Sensors**	
28.1	*The sensors I am going to attach to you are:* 1. an **electrocardiogram, or EKG** which will go on your upper torso, beneath your clothing 2. a **respiration belt**, which will go on your ribcage, above your clothing 3. a **blood pressure cuff** on your left arm.	Complete? ☐
Notes		

29.0	**EKG electrode placement**	
29.1	Attach the EKG electrodes to the subject:	
	29.1.1	*First, I am going to attach EKG electrodes on you like this picture.* *[Gestures to picture on wall.]*	

	29.1.2	Steps:
		
		1. *Please* **stand up**.
		
		2. Put on a pair of **disposable gloves**.
		
		3. Clean oils on skin with **alcohol wipes**
		
		    (a) **Boniest part of the right shoulder.**
		    (b) **Left lowest rib, 1 inch to the left of your nipple.**
		    (c) **Area below right rib, 1 inch to the right of your nipple and 2 inches below the left electrode position.**
		
		4. Let the **alcohol dry a bit**.
		
		5. Attach the electrodes[14].
		
		6. *You can sit down now.*
			Complete?
			
			☐

	29.1.3	*Are you doing ok?*	
Notes		

---

[14]If the electrodes fall off, clean the skin again and use a new electrode. Note that electrode patches can dry out and become less sticky if exposed to air.

30.0		**Respiration Belt**	
30.1		Attach the respiration belt to the subject:	
	30.1.1	*Next, we need to attach the respiration belt like this picture.* [*Gestures to picture on wall.*]	
	30.1.2	Steps: 1. *This respiration belt goes* **over your clothes** *and* **high around your chest**. 2. *It needs to be* **right side up** *and* **snug**. 3. *If it isn't snug, we won't get any readings but we want you to be comfortable too.* 4. [Experimenter wraps the respiration belt around the subject.] 5. *You should only feel constrained when you take a deep breath.* 6. *Can you take a deep breath now?*[15]	Complete? ☐
	30.1.3	*Are you doing ok?*	
Notes			

[15]Once on, the belt should fit snugly. It should feel modestly constraining and should produce a sense of discomfort if you take a very deep breath.

31.0	**EKG lead placement**	
31.1	Attach the EKG leads to the subject:	
	31.1.1	*Next, we're going to attach these EKG leads to the electrodes like this picture. [Gestures to picture on wall.]*	
	31.1.2	Steps: 1. *I'll give you a lead and you'll snap them on like this.* [Demonstrates how to snap on a spare electrode.] 2. The **white lead** goes on the **right shoulder electrode.** 3. The **black lead** goes on the **left rib electrode.** 4. The **green lead** goes on the **right torso electrode.** 5. *Let me check to see if they are* **secure.** 6. **[Gently tugs at leads.]** 7. *Can you please sit as though you were going to type?*	Complete? ☐
	31.1.3	*Are you doing ok?*	
Notes		

32.0	**Blood Pressure Cuff**	
32.1	Attach the blood pressure cuff to the subject. *Now, I am going to attach the blood pressure cuff on you again.*	

32.1.1	Steps:
	
	1. *Which hand do you usually use for the mouse?*
	
	2. Have subject remove clothing or move fabric for the **non-mouse arm**[16].
	
	3. Place the cuff on the upper arm. Make sure the:
	
	    (a) **air is squeezed out of the cuff**
	    (b) **muscles are not tense**
	    (c) BP cuff has **some overlap of velcro**
	    (d) **BP cuff artery arrow is pointed at the pulse** (it's slightly off-center and about 1" from the elbow crease)
	    (e) the **cuff index line falls within the range markings**. If it's too small/big, replace the cuff with the appropriate size.
	
	4. Then, check to ensure the:
	
	    (a) **tightness of the cuff allows 1-2 fingers to slip in**
	    (b) **velcro attachments are secure**
	    (c) **hose connected to the cuff is not kinked or warped**
		Complete? ☐

32.2	Provide the subject with instructions for the blood pressure cuff:	
32.2.1	1. *This BP cuff will stay on your arm for the whole experiment.* 2. *Readings will be taken **automatically every 5 minutes**.* 3. **Don't overly flex your arm** *or* **lift it above your shoulder**. 4. *Your arm should stay on the armrest or at your side.* 5. *Let me know if you feel the cuff is slipping off.*	Complete? ☐
32.3	Use strips of masking tape so **all wires are out of the way.**	Complete? ☐
Notes		

---

[16]If the cuff is going on the right arm, make sure the hose runs over the subject's lap and is comfortable.

33.0	**Confirming Readings**		
33.1	Start up LabChart:		
	33.1.1	*Now I'm going check to the readings for the EKG and respiration belt.*	
	33.1.2	On Labmon, double click on the desktop shortcut: **Stress_expt**	Complete? ☐
	33.1.3	Ensure this miniwindow is visible: **Preset Comment Function Keys**	Complete? ☐
	33.1.4	Click the button: **Start**	Complete? ☐
33.2	The experimenter will **confirm that appropriate readings are coming in**.		Complete? ☐
	33.2.1	Adjust lab chart so there are high spikes above and low spikes below for the EKG and Respiration readings.	Complete? ☐
	33.2.2	The incoming readings should be similar to the reference EKG and Respiration example on the wall.	
	33.2.3	IF RESPIRATION BELT READINGS ARE INAPPROPRIATE:[17] 1. **Ensure that all connections are firm.** 2. **Ensure all the equipment is powered on.** 3. **Ask the subject to place his/her hand over the sensor.** 4. **Ensure that the sensor is positioned correctly–facing up in the center of the chest.** 5. **Ask the subject to confirm that the belt is tight against his/her stomach.** 6. **Ask the subject to take a deep breath and hold it in for a few seconds.** 7. **Check that the channel range is appropriate.**[18]	

---

[17]Respiration belt readings should be rolling hills. The readings should automatically adjust after a five seconds.

[18]If the reading is still inadequate, consult the Troubleshooting section in the operations manual. The range can be adjusted by clicking on the **drop-down arrow** and selecting one of the options.

	33.2.4	IF EKG READINGS ARE INAPPROPRIATE **(signals are upside down or not the expected QRS wave)**[19]: 1. **Ensure that all connections are firm.** 2. **Ensure all the equipment is powered on.** 3. **Loosely pull on the leads; the wires should not come out if the leads are properly clipped.** 4. **If the electrodes are improperly placed or the leads are not properly clipped, ask the subject to remedy the problem.**[20]	
Notes		

---

[19]The lab chart should have high spikes above and low spikes below.

[20]If these steps do not resolve the issue, consult the Troubleshooting section in the operations manual.

34.0	**1st neutral induction**		
34.1	Guide the subject through the first rest period:		
	34.1.1	**Turn off the fluorescent lights**.	Complete? ☐
	34.1.2	Instructions: 1. *We will now begin a 30-minute rest period, where you should sit back and relax.* 2. *To help you relax, I'd like you to watch a video of peaceful underwater scenes and perform a simple task.* 3. *I'd like you to note the different categories of animals that appear in the video.* 4. *Use very general categories like fish and birds as though you were describing the animals to a small child.* 5. *Please write the different animal categories down on this piece of paper. You only need to write down an animal once.* [Hand the subject pen and paper.] 6. *Please remember to* **keep your feet flat on the floor**. 7. *Is your BP cuff still secure?* 8. *I'll start the automatic blood pressure readings now. They will happen roughly every 5 minutes.*	Complete? ☐
	34.1.3	**Turn the Critikon on.**	Complete? ☐
	34.1.4	Press the **Cycle button** on the Critikon until it reads **5 minutes**.[21]	Complete? ☐
	34.1.5	**Press F5 in LabChart** to insert a comment indicating the start of the blood pressure readings.	(**F5**)
	34.1.6	**Press F1 in LabChart** to insert a comment indicating the start of the rest period. [22]	(**F1**)
.	34.1.7	Start the first video on the experimenter's laptop with: **run-video1.bat**	Complete? ☐
	34.1.8	The experimenter will **start the timer.** START TIME: _____	Complete? ☐
	34.1.9	During the rest period, **ensure that the subject does not fall asleep**. If necessary, wake the subject by saying their name or tap their shoulder.	

---

[21]If the alarm sounds, press the yellow button twice to silence the alarm. Resume blood pressure readings by hitting the cycle button and pressing the button to take a blood pressure reading.

[22]This can be also be done by clicking F1 in the **Preset Comments Window** or by entering text into the text box at the top of the LabChart window and clicking the **Add button**

34.2		**Check that the subject's feet are flat on the floor**. If they move their feet, remind them to keep them flat.	Complete? ☐
	34.2.1	After 30 minutes, **turn off the timer**.	Complete? ☐
	34.2.2	**Press F2 in LabChart** to insert a comment indicating the end of the rest period.	(F2)
	34.2.3	**Turn off the Critikon**.	Complete? ☐
	34.2.4	Stop the video by **pressing Escape and then clicking on the red X in the corner**.	Complete? ☐
	34.2.5	*The rest period is now over.* Take **the paper** from the subject and set it aside.	Complete? ☐
34.3		Administer **NASA-TLX** and **VAS**. 1. Give forms to subject. *Please fill out these forms.* 2. Check to see if **subject filled out both forms**. 3. Place the **NASA-TLX** and **VAS** in the manila folder.	Complete? ☐
Notes			

35.0	**1st neutral typing sample**	
35.1	Prepare for the 1st neutral typing sample:	
	35.1.1	Instructions: 1. *We are now going to start the first typing sample.* 2. *Let me* **start the video cameras** *and* **the typing software**. 3. *Please move the mouse out of the way; you won't need it now.* 4. *Go ahead and adjust the keyboard so that you are in a comfortable typing position.* 5. *Then, place your hands as if you were about to start typing.*	Complete? ☐
35.2	Start up the cameras in command window using: **start-cameras.bat** **<subject-number>**	Complete? ☐
	35.2.1	The experimenter will check the videos on Photomon are displaying: 1. **Right camera** - **keyboard** 2. **Top camera** - **keyboard**, **labels** 3. **Left camera** - **keyboard** 4. **Face camera** - **subject's entire face, green background** 5. Make sure the **Mouse is out of the way**.	Complete? ☐
	35.2.2	*Could you please wiggle your fingers a bit? Ok, that looks good.*	Complete? ☐
35.3	Start MTP in the command window of the experimenter's laptop: **run-cog-n1.bat** **<subject-number>**	Complete? ☐
	35.3.1	**Move cursor** to the corner of the screen.	Complete? ☐
	35.3.2	Start the subject on the 1st typing sample: 1. *This task will have* **80 repetitions** *of the same phrase as before.* 2. *Please remember* **not to talk or pause in the middle of a phrase**. 3. *Also,* **keep your feet flat on the floor** *while you type. Ok?* 4. *Please start typing now.*	Complete? ☐
	35.3.3	**Press F3 in LabChart** to insert a comment indicating the start of a typing session.	**F3**
	35.3.4	The experimenter will note the start time. **TIME:** _____	Complete? ☐

35.4	As the subject types, take notes on typing style.		**Complete?** ☐
	35.4.1	Wrist support:	
	35.4.2	Isolated digits:	
	35.4.3	Typing force:	
	35.4.4	Other:	
35.5	**Check LabChart** to make sure the sensor readings are being taken.		**Complete?** ☐
35.6	**Check that the subject's feet are flat on the floor**. If they move their feet, remind them to keep them flat.		**Complete?** ☐
	35.6.1	**Press F4 in LabChart** to insert a comment indicating the end of a typing session.	**F4**
	35.6.2	The experimenter will note the stop time. **TIME:** _____	**Complete?** ☐
	35.6.3	Close MTP and press **'q'** in the command window.	**Complete?** ☐
35.7	Administer **NASA-TLX** and **VAS**. 1. Give forms to subject. *Please fill out these forms.* 2. Check to see if **subject filled out both forms**. 3. Place the **NASA-TLX** and **VAS** in the manila folder.		**Complete?** ☐
Notes			

36.0	**Purple exercise**	
36.1	Set up Purple:	
	36.1.1	Instructions: 1. *Next up, you will be doing a longer version of the multi-tasking exercise.* 2. *Please move the keyboard out of the way and put the* **Mouse** *in a comfortable position.* 3. *Then, I can check if the video looks good.*	Complete? ☐
	36.1.2	The experimenter will check the videos on Photomon are displaying: 1. **Right camera** - **keyboard** 2. **Top camera** - **keyboard**, **labels** 3. **Left camera** - **keyboard** 4. **Face camera** - **subject's entire face, green background** 5. *Ok, that looks good. Let me start up Purple.*	Complete? ☐
	36.1.3	Start up the Purple software in the command window of the experimenter's laptop with: 1. **run-purple.bat** 2. Enter the <**subject-number**> and **Task.cfg**.	Complete? ☐
36.2	Purple instructions and BP setup:	
	36.2.1	1. *You must aim to be fast and accurate on* **ALL** *of the tasks.* 2. **$30 of your compensation depends on how well you perform on this task.** 3. *I'll be evaluating your performance based on your score and will also check to see whether you are attending to all of the tasks.* 4. *Please remember to keep* **both your feet flat on the floor** *while you use the software.* 5. *Do you have any questions?* 6. *Ok. Let me start up the BP readings, then you can start.*	Complete? ☐
	36.2.2	**Turn on the Critikon.**	Complete? ☐

	36.2.3	Press the **Cycle button** on the Critikon until it reads **5 minutes**. [23]	Complete? ☐
	36.2.4	**Press F5 in LabChart** to insert a comment indicating the start of the blood pressure readings.	(**F5**)
	36.2.5	Press **F6** to mark the start of the Purple exercise.	(**F6**)
	36.2.6	*Please start now.*	Complete? ☐
	36.2.7	**START TIME:** _____	
	36.2.8	The experimenter will **start the timer**.	Complete? ☐
	36.2.9	**Check LabChart** to see whether the sensors are working.	Complete? ☐
	36.2.10	Demonstrate active tracking of the subjects performance by: 1. **Standing close** to the subject 2. Looking at their screen	
	36.2.11	Monitor the subject for signs of engagement: 1. Subject should be averaging at least 2 clicks per second 2. Subject should be looking at the screen at all times 3. Subject's eyes should be looking at the different quadrants on the screen 4. Subject's facial expression should be focused 5. Subject should seem occasionally frustrated (e.g., frowning, shaking head, making discontented noises), especially after making mistakes or when you criticize his/her performance 6. Subject's posture should be engaged – generally upright and leaning towards the screen 7. Subject should NOT be yawning 8. Subject should NOT be slouching	Complete? ☐

---

[23]If the alarm sounds, **press the yellow button twice** to silence the alarm. Resume blood pressure readings by hitting the cycle button and pressing the button to take a blood pressure reading.

	36.2.12	If you observe that the subject is NOT engaged, criticize their performance:
		
		   1. *Make sure you are focusing on the task.*
		
		   2. *You seem distracted, please refocus on the task.*
		
		   3. *Please concentrate on the task. Remember, you must stay focused and perform well to get the \$30.*
		
		   4. *You are not working fast enough.*
		
		   5. *You are making too many errors.*
			Complete? ☐

	36.2.13	At the **5 minute mark**, the experimenter will administer social evaluation:
		
		   1. Lean in and look at the subject's computer screen.
		
		   2. *Could you work faster? Most subjects have over <score + 200/500> points by now.*[24]
			Complete? ☐

	36.2.14	At the **10 minute mark**, the experimenter will administer social evaluation:
		
		   1. Lean in and look at the subject's computer screen.
		
		   2. *You really need to work faster to earn the full \$30.*
			Complete? ☐

	36.2.15	At the **15 minute mark**, Purple will stop automatically.	
	36.2.16	**Press F7 in LabChart** to mark the end of the Purple exercise.	(F7)
	36.2.17	**Turn off the Critikon.**	Complete? ☐

36.3		Administer **NASA-TLX** and **VAS**.
		
		   1. Give forms to subject. *Please fill out these forms.*
		
		   2. Check to see if **subject filled out both forms**.
		
		   3. Place the **NASA-TLX** and **VAS** in the manila folder.
			Complete? ☐

Notes			

---

[24]If the subject talks back or otherwise indicates displeasure at the comment, ask them to please focus on the task.

37.0	**Cognitive load typing sample**	
37.1	Set up for the cognitive load typing sample:	
	37.1.1	Instructions: 1. *We are now going to start the second typing sample.* 2. *Let me* **check the video cameras** *and* **start the typing software**. 3. *Please move the mouse out of the way; you won't need it now.* 4. *Go ahead and adjust the keyboard so that you are in a comfortable typing position.* 5. *Then, place your hands as if you were about to start typing.*	Complete? ☐
	37.1.2	The experimenter will check the videos on Photomon are displaying: 1. **Right camera** - **keyboard** 2. **Top camera** - **keyboard**, **labels** 3. **Left camera** - **keyboard** 4. **Face camera** - **subject's entire face, green background** 5. Make sure the **Mouse is out of the way**.	Complete? ☐
	37.1.3	*Could you please wiggle your fingers a bit? Ok, that looks good.*	Complete? ☐
37.2	Start MTP in the command window of the experimenter's laptop: **run-cog-cog.bat** <**subject-number**>	Complete? ☐
	37.2.1	**Move cursor** to the corner of the screen.	Complete? ☐
	37.2.2	Start the subject on the 2nd typing sample: 1. *This task will be just like the last one.* 2. *Please remember* **not to talk or pause in the middle of a phrase**. 3. *Also,* **keep your feet flat on the floor** *while you type. Ok?* 4. *Please start typing now.*	Complete? ☐
	37.2.3	**Press F3 in LabChart** to insert a comment indicating the start of a typing session.	(F3)
	37.2.4	The experimenter will note the start time. **TIME:** _____	Complete? ☐
37.3	The experimenter **Stands up**.	Complete? ☐

			Complete?
37.4		As the subject types, take notes on typing style.	☐
	37.4.1	Wrist support:	
	37.4.2	Isolated digits:	
	37.4.3	Typing force:	
	37.4.4	Other:	
37.5		**Check LabChart** to make sure the sensor readings are being taken.	Complete? ☐
37.6		**Check that the subject's feet are flat on the floor**. If they move their feet, remind them to keep them flat.	Complete? ☐
	37.6.1	**Press F4 in LabChart** to insert a comment indicating the end of a typing session.	(F4)
	37.6.2	The experimenter will note the stop time. **TIME:** _____	Complete? ☐
	37.6.3	Close MTP and press **'q'** in the command window.	Complete? ☐
37.7		Administer **NASA-TLX** and **VAS**. 1. Give forms to subject. *Please fill out these forms.* 2. Check to see if **subject filled out both forms**. 3. Place the **NASA-TLX** and **VAS** in the manila folder.	Complete? ☐
Notes			

38.0	**2nd neutral induction**	
38.1	Guide the subject through the second rest period:	
38.1.1	Instructions: 1. Tell subject that they did well on Purple and will get the full $30. Also tell them that they are very likely to get the full $60 if they continue to do well for this last rest period and last typing sample. 2. *Ok, we'll now begin the second 15-minute rest period.* 3. *Please rest and relax, while you watch another video of peaceful underwater scenes and do a simple belly breathing exercise.* 4. *In this belly breathing exercise, we want you to take long, deep breaths that cause your stomach to move in and out instead of just your chest.* 5. *While you do this, please focus on relaxing as much as possible (without falling asleep).* 6. *You need to* **keep your head still** *so your face stays in the video. Also, please* **don't touch your face***.* 7. *And remember to* **keep your feet flat on the floor***.* 8. *Is your BP cuff still secure?*	Complete? ☐
38.1.2	**Turn the Critikon on.**	Complete? ☐
38.1.3	Press the **Cycle button** on the Critikon until it reads **5 minutes**.[25]	Complete? ☐
38.1.4	**Press F5 in LabChart** to insert a comment indicating the start of the blood pressure readings.	(F5)
38.1.5	**Press F1 in LabChart** to insert a comment indicating the start of the rest period. [26]	(F1)
38.1.6	Start the video on the experimenter's laptop with: **run-video2.bat**	Complete? ☐
38.1.7	The experimenter will **start the timer.** START TIME: _____	Complete? ☐
38.1.8	During the rest period, **ensure that the subject does not fall asleep**. If necessary, wake the subject by saying their name or or tap their shoulder.	

---

[25]If the alarm sounds, press the yellow button twice to silence the alarm. Resume blood pressure readings by hitting the cycle button and pressing the button to take a blood pressure reading.

[26]This can be also be done by clicking F1 in the **Preset Comments Window** or by entering text into the text box at the top of the LabChart window and clicking the **Add button**

38.2		**Check that the subject's feet are flat on the floor**. If they move their feet, remind them to keep them flat.	Complete? ☐
	38.2.1	After 15 minutes, **turn off the timer**.	Complete? ☐
	38.2.2	**Press F2 in LabChart** to insert a comment indicating the end of the rest period.	(F2)
	38.2.3	**Turn off the Critikon**.	Complete? ☐
	38.2.4	Stop the video by **pressing Escape and then clicking on the red X in the corner**.	Complete? ☐
	38.2.5	*The rest period is now over.*	Complete? ☐
38.3		Administer **NASA-TLX** and **VAS**. 1. Give forms to subject. *Please fill out these forms.* 2. Check to see if **subject filled out both forms**. 3. Place the **NASA-TLX** and **VAS** in the manila folder.	Complete? ☐
Notes			

39.0	**2nd neutral typing sample**	
39.1	Set up for the 2nd neutral typing sample:	
	39.1.1	Instructions: 1. *We are now going to start the last typing sample.* 2. *Let me* **check the video cameras** *and* **start the typing software**. 3. *Go ahead and adjust the keyboard so that you are in a comfortable typing position.* 4. *Then, place your hands as if you were about to start typing.*	Complete? ☐
	39.1.2	The experimenter will check the videos on Photomon are displaying: 1. **Right camera** - **keyboard** 2. **Top camera** - **keyboard**, **labels** 3. **Left camera** - **keyboard** 4. **Face camera** - **subject's entire face, green background** 5. Make sure the **Mouse is out of the way**.	Complete? ☐
	39.1.3	*Could you please wiggle your fingers a bit? Ok, that looks good.*	Complete? ☐
39.2	Start MTP in the command window of the experimenter's laptop: **run-cog-n2.bat** <**subject-number**>	Complete? ☐
	39.2.1	**Move cursor** to the corner of the screen.	Complete? ☐
	39.2.2	Start the subject on the 2nd neutral typing sample: 1. *This task will be just like the last one.* 2. *Please remember* **not to talk or pause in the middle of a phrase**. 3. *Also,* **keep your feet flat on the floor** *while you type. Ok?* 4. *Please start typing now.*	Complete? ☐
	39.2.3	**Press F3 in LabChart** to insert a comment indicating the start of a typing session.	(**F3**)
	39.2.4	The experimenter will note the start time. **TIME:** _____	Complete? ☐
39.3	As the subject types, take notes on typing style.	Complete? ☐
	39.3.1	Wrist support:	

	39.3.2	Isolated digits:	
	39.3.3	Typing force:	
	39.3.4	Other:	
39.4	**Check LabChart** to make sure the sensor readings are being taken.		Complete? ☐
39.5	**Check that the subject's feet are flat on the floor**. If they move their feet, remind them to keep them flat.		Complete? ☐
	39.5.1	**Press F4 in LabChart** to insert a comment indicating the end of a typing session.	(**F4**)
	39.5.2	The experimenter will note the stop time. **TIME:** _____	Complete? ☐
	39.5.3	Close MTP and press **'q'** in the command window.	Complete? ☐
39.6	Administer **NASA-TLX** and **VAS**. 1. Give forms to subject. *Please fill out these forms.* 2. Check to see if **subject filled out both forms**. 3. Place the **NASA-TLX** and **VAS** in the manila folder.		Complete? ☐
Notes			

40.0	**Save readings and remove sensors**			
40.1	Save readings:			
	40.1.1	Stop the video on Photomon with: **stop-cameras.bat**		Complete? ☐
	40.1.2	**Stop the LabChart recording** by pressing: **Stop**		Complete? ☐
	40.1.3	**Save the LabChart reading** 1. Select **File** -> **Save As...** 2. Save the LabChart reading in the Experiments folder with the <**subject number**>. 3. **Minimize the Labchart recording** and check that the file exists in the Experiments folder.		Complete? ☐
	40.1.4	**Print out the blood pressure readings** 1. **Turn the Critikon on.** 2. Press **the History button** 3. Press **the Print button**. Double check that there are readings every 5 minutes from the start of the experiment to the end. 4. **Staple** printout to manila folder. 5. **Turn the Critikon off.**		Complete? ☐
40.2	Remove the sensors from the subject:			
	40.2.1	Turn on the room lights.		Complete? ☐
	40.2.2	*We're done with the experiment so let's remove everything carefully.* 1. Detach blood pressure cuff. 2. Detach respiration belt. 3. *Could you please carefully peel off the electrode pads with the leads still attached to them and give them to me? Thanks.* 4. *Do you need an* **alcohol swab** *for any residue?* Gestures to alcohol swabs.		Complete? ☐
Notes				

41.0	**Debriefing**	
41.1	Debrief the subject:	
	41.1.1	*This concludes the experiment. Thank you for your participation.*	
	41.1.2	IF THE SUBJECT WAS ENGAGED: 1. *You performed well during the experiment, so you get the full 60 dollars.* 2. The experimenter provides **60 dollars** to the subject. 3. *Please sign next to your name on the* **Subject Payment Record form***.*	Complete? ☐
	41.1.3	IF THE SUBJECT WAS *NOT* ENGAGED 1. *Unfortunately, you did not perform well enough during the experiment to earn the additional bonus.* 2. The experimenter provides **10 dollars** to the subject. 3. *Please sign next to your name on the* **Subject Payment Record form***.*	
	41.1.4	*I have a few questions before you go:* 1. *Did you feel like you* **zoned out** *a bit during the typing tasks?* 2. *Did you experience cognitive load during Purple?* 3. *Did you find the videos calming or boring?* 4. *Did you prefer the animals or the breathing task?* Response:	Complete? ☐
	41.1.5	*Thank you for your time today. Please make sure* **you collect all your belongings** *before you go.* Gestures to box and coat rack.	Complete? ☐
	41.1.6	The subject will depart.	
	41.1.7	Experimenter will take down the **Experiment in Progress** sign, **turn up the phone ringer**, and **remove the desk label**.	Complete? ☐
Notes		

42.0	**Clean-up**		
42.1	Save and backup the collected data.		
	42.1.1	Copy all four of the "Cognitive Loading" folders from the experimenter's laptop onto the USB stick and place it next to the laptop. Shing-hon will be in charge of taking the data and storing it on Coolmon.	Complete? ☐
	42.1.2	Transfer the hand photo data from the hand-photo laptop to Coolmon.	Complete? ☐
	42.1.3	Video data is automatically stored on Photomon. Shing-hon will be in charge of backing up this data to Coolmon.	
42.2	**Turn off PowerLab**.		Complete? ☐
42.3	**Turn off the LED lights.**		Complete? ☐
42.4	Use alcohol wipes to clean the: 1. Blood pressure cuff 2. EKG leads		Complete? ☐
42.5	Still camera clean-up. 1. Ensure the **still camera and the remote are off**. 2. Unplug the still camera remote and plug in the charger. **BE CAREFUL! The mini-USB is very delicate and prone to bending!** 3. Cover the still camera with the cloth.		Complete? ☐
42.6	If you haven't done so already, **Copy the initials and date of birth** from the demographic form to all the other forms.		Complete? ☐
42.7	Ensure that the manila folder contains:		Complete? ☐
	42.7.1	1 copy of the consent form, signed and dated by both the subject and the experimenter.	Complete? ☐
	42.7.2	1 copy of the demographic survey	Complete? ☐
	42.7.3	1 copy of the long-form STAI Y-2.	Complete? ☐
	42.7.4	1 copy of the PSS.	Complete? ☐
	42.7.5	6 copies of the NASA-TLX.	Complete? ☐
	42.7.6	6 copies of the VAS.	Complete? ☐

			Complete?
	42.7.7	1 blood pressure reading printout.	☐
42.8		Place this protocol checklist into the manila folder and place the manila folder into the file cabinet.	Complete? ☐
42.9		If you haven't done so already, **Print out the blood pressure readings** 1. **Turn the Critikon on.** 2. Press **the History button** 3. Press **the Print button**. Double check that there are readings every 5 minutes from the start of the experiment to the end. 4. **Staple** printout to manila folder. 5. **Turn the Critikon off.**	Complete? ☐
42.10		If you have not done so already, **Save the LabChart reading** 1. Select **File -> Save As...** 2. Save the LabChart reading in the Experiments folder with the <**subject number**>. 3. **Minimize the Labchart recording** and check that the file exists in the Experiments folder.	Complete? ☐
Notes			

A.8 Operations Manual

Our experiment involves numerous pieces of equipment, each essential to the successful conduct of the experiment. It is obviously critical that the experimenter knows how each piece of equipment works, and we decided that a written operations manual would assist our experimenter in achieving that understanding. Moreover, we noted at the outset of the design of the experiment that a foreseeable risk in executing our experiment is that one of these pieces of equipment might breakdown prior to or while running a subject. Many of these breakdowns are fairly easy to remedy and repairs could be effected by the experimenter in short order. These common breakdowns and the repairs required to remedy them were also added into the operations manual. As with the protocol, we reproduce here the full operations manual document both for the sake of transparency and in the hope that others might find it helpful in their own research.

Stress Experiment – Operations Manual

Shing-hon Lau

2-17-2016 at 14:30

Contents

1 Machine access

There are three machines used in the stress experiment. The first is the laptop, which is used for all keystroke experiments. The laptop is not password protected. The second machine is Photomon, which runs the video-capture software. The third machine is Labmon, which runs the physiological software. The username on both machines is *administrator* and the password is *adm∼* followed by the asset tag code, which is 020297 for Photomon and 020295 for Labmon. The asset tag code can be found on the top of both machines.

2 Video

There are three primary reasons that we collect video in the stress experiment. First, the video enables us to associate any interesting typing timings or physiological readings with subject behavior. Second, the video allows us to subjectively determine whether the subject is stressed. Third, the face video will be provided to researchers at Pitt (SHL: What researchers? What are they doing?).

The video cameras setup consists of four different cameras: one focused on the face, one focused on the keyboard from above, one focused on the keyboard from the left side, and one focused on the keyboard from the right side. All four cameras are connected to a single video machine. Each camera's video is captured independently, into its own file, at a resolution of 720p (1280x720) using the Open Broadcaster Software (OBS) – an open source video capturing software.

2.1 Operating Procedures

Starting all the cameras at once

1. If no command window is open, double click on the `Command Window` shortcut on the Desktop.

2. Ensure that the path in the command window reads `C:\Users\Administrator\Desktop`. If it does not, close the window and return to step 1.

3. At the command line, type `start-cameras.bat <subject number>` and press `Return`, replacing `<subject number>` with the subject number for the current subject.

4. All cameras will appear on the screen, with each camera in its own window. Note: Cameras start recording as soon as the image appears on the screen.

Stopping all the cameras at once

1. If no command window is open, double click on the `Command Window` shortcut on the Desktop.

2. Ensure that the path in the command window reads `C:\Users\Administrator\Desktop`. If it does not, close the window and return to step 1.

3. At the command line, type `stop_cameras.bat` and press `Return`. No subject number is required.

4. All windows with a camera image will close. Note: Cameras stop recording as soon as they disappear from the screen.

Restarting a camera

1. Click on "Stop Recording".

2. Wait 3 seconds.

3. Click on "Start Recording".

4. If this does not fix the problem, continue to the remaining steps.

5. Close the OBS window containing the camera that has failed.

6. If no command window is open, double click on the `Command Window` shortcut on the Desktop.

7. Ensure that the path in the command window reads `C:\Users\Administrator\Desktop`. If it does not, close the window and return to step 1.

8. To start the left camera, type `start-left-camera.bat <subject number>` and press `Return`, replacing `<subject number>` with the subject number for the current subject.

9. To start the right camera, type `start-right-camera.bat <subject number>` and press `Return`, replacing `<subject number>` with the subject number for the current subject.

10. To start the top camera, type `start-top-camera.bat <subject number>` and press `Return`, replacing `<subject number>` with the subject number for the current subject.

11. To start the face camera, type `start-face-camera.bat <subject number>` and press `Return`, replacing `<subject number>` with the subject number for the current subject.

12. The missing camera(s) will now appear on the screen, with each camera in its own window. Note: Cameras start recording as soon as the image appears on the screen.

2.2 Troubleshooting

There is an issue with OBS or one of the cameras.

|	When this happens	What to do
1.	The batch file gives an error	Ensure that all of the cameras are plugged in (three in back and one in front).
		If it still does not work, attempt to start the cameras individually.
		If it still does not work, stop all cameras and start them all again.
		Make a note of the issue in the checklist.
2.	One or more of the cameras are black OR have frozen.	Restart all cameras by stopping and then starting them.
		Check to see that the cameras are not affected by the camera bug (see Troubleshooting item number 6, below).
		If it still does not work and the experiment has not yet started, ensure that the cameras are plugged in (three in back and one in front).
		If it still does not work and the experiment has not yet started, unplug the cameras and then plug them back in.
		If it still does not work and the experiment has not yet started, restart Photomon.
		Make a note of the issue in the checklist and proceed without the affected cameras.
3.	The video is distorted.	Restart all cameras by stopping and then starting them.
		If it still does not work, ensure that the cameras are firmly plugged in (three in back and one in front).
		If the issue persists and the experiment has not yet started, restart the video machine.
		If the issue persists but the experiment has already started, proceed without video.
		Make a note of the issue in the checklist.
4.	OBS has stopped responding.	Close all the OBS windows by clicking on the red X.
		If OBS refuses to close, run `stop_cameras.bat` and then restart the cameras.
		If the issue persists and the experiment has not yet started, restart the video machine.
		If the issue persists but the experiment has already started, proceed without video.
		Make a note of the issue in the checklist.
5.	The cameras do not have the proper field of view.	Slightly reposition the subject or keyboard to ensure a proper field of view.
		Make slight adjustments to the cameras to ensure a proper field of view.
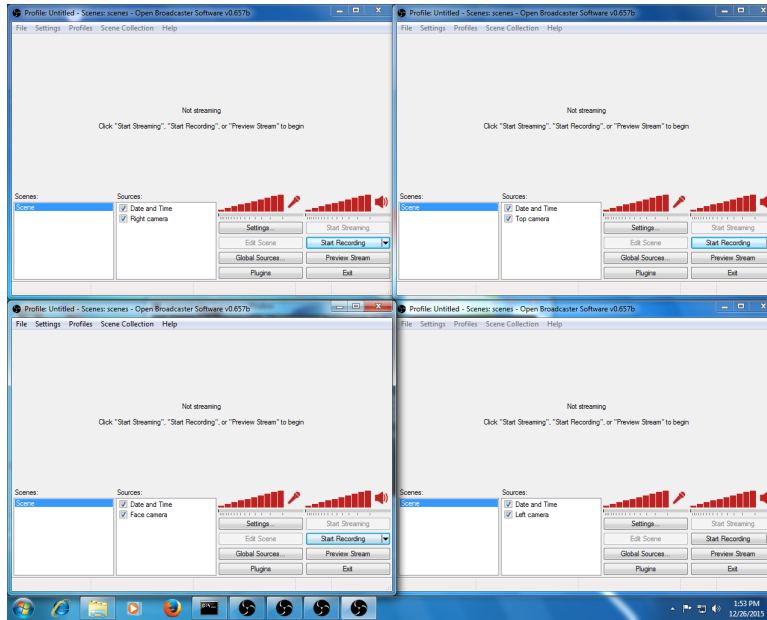
Figure 1: OBS view with all cameras turned off.

6. A camera is black due to the camera bug.	Identify the video camera that is black (should be one of the 4 quadrants as in Fig. 1).
	Right click on the camera name in that window and select properties (as in Fig. 2).
	From the properties menu, select the first "Microsoft LifeCam Studio" camera that appears (as in Fig. 3).
	Hit OK on the properties menu and then click on "Preview Stream" in the main window (Fig. 1).
	If the issue is not yet fixed, repeat the above steps with the second, third, and fourth "Microsoft LifeCam Studio" cameras.
	When the issue is fixed, click "Stop Preview" and click "Start Recording".
	If the issue is still not fixed, proceed with the remaining troubleshooting steps in Step 2.
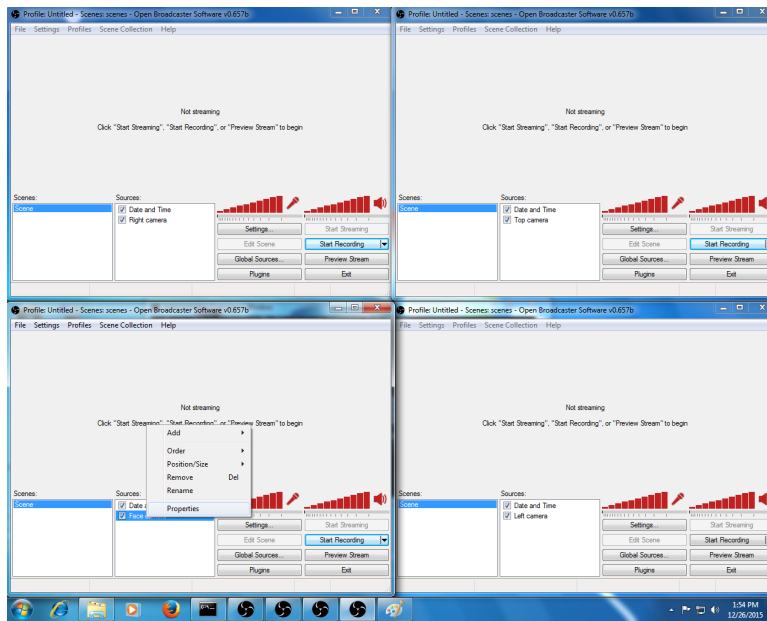
Figure 2: Selecting the properties of the affected camera.
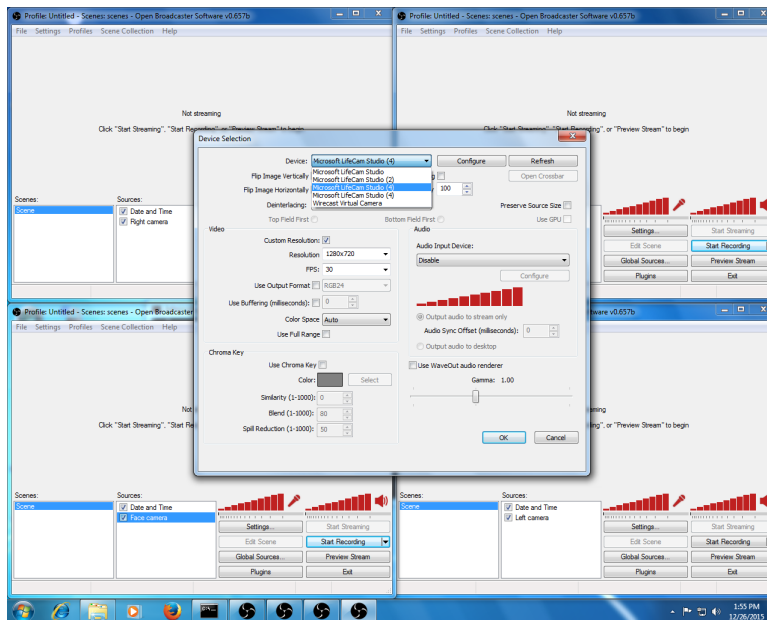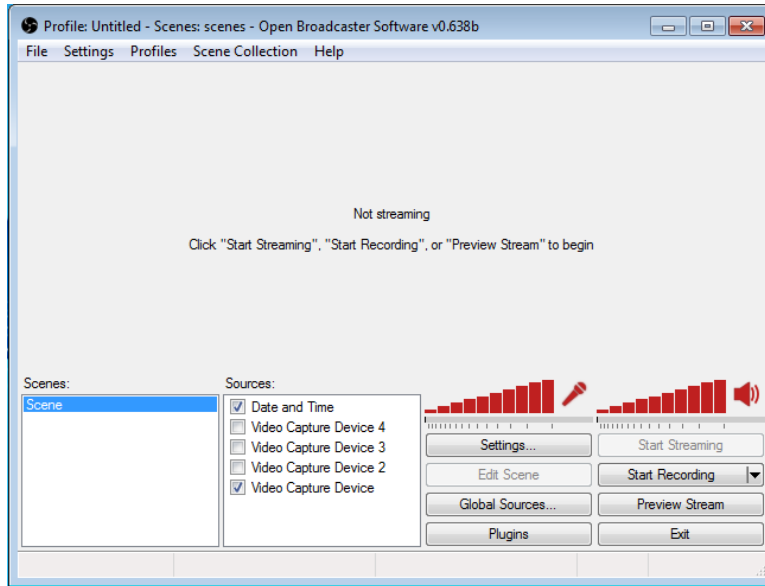

Figure 3: Reassigning the affected camera.

Figure 4: OBS's main window.

2.3 Technical Details

2.3.1 Cameras

Each of the four cameras has the same make and model: Microsoft LifeCam Studio (`http://www.microsoft.com/hardware/en-us/p/lifecam-studio`). The most important characteristic of this camera is the 1080p resolution. Additionally, the camera comes on a mount, permitting it to be attached to various extension and mounting equipment. This permits us to get the best possible camera angle for each of our cameras. The camera is a USB 2.0 camera.

After we had purchased these cameras, we considered upgrading to a Logitech C920 (`http://www.logitech.com/en-us/product/hd-pro-webcam-c920`), which has a wider field of view. However, we opted not to because the Microsoft LifeCam Studio cameras were already purchased.

2.3.2 Open Broadcaster Software (OBS)

OBS (`https://obsproject.com/`)is a piece of software intended primarily to facilitate the live-streaming of events. It has the ability to take in a large number of sources, including cameras, text, images, and audio. We are using this software for one of its secondary features, which is the ability to record a broadcast to disk. The software is highly customizable, permitting nearly every aspect of the recording to be customized.

Figure 4 shows the main window for OBS. The bottom left of the screen contains the scene selection; scenes are defined pre-sets for which cameras, text overlays, and graphical overlays are recorded. For our purposes, there will only ever be a single scene. In the bottom middle of the screen is the list of sources used in the scene. Sources include cameras, text overlays, and graphical overlays. It is also possible to record programs on the machine. The bottom right of the screen contains buttons to configure OBS and to start/stop recording, start/stop the video preview, and exit OBS. The upper half of the window displays the video being captured or a message stating that the recording has not started yet. Once preview mode has been turned on, it is possible to edit the scene by clicking the "Edit scene" button. This allows drag-and-drop editing of the elements
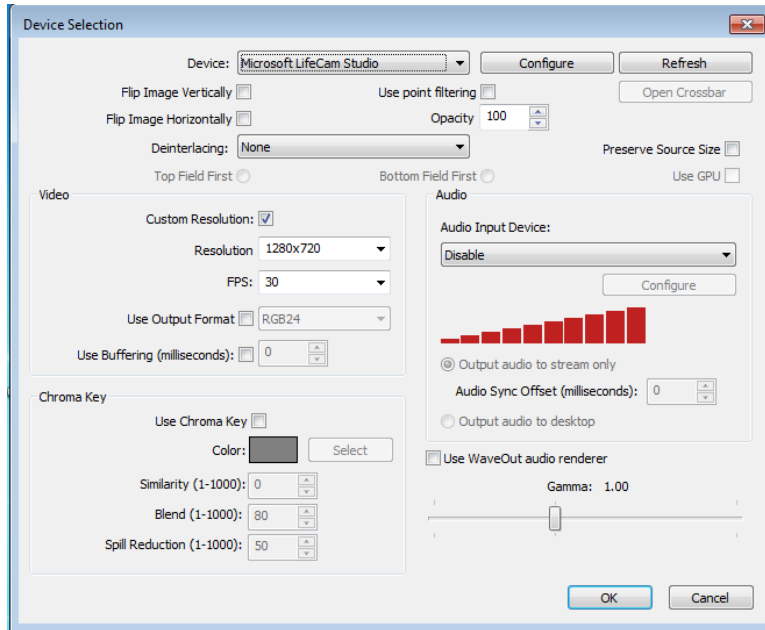
8

Figure 5: Camera settings window in OBS.

in the scene.

To add a new camera (or any other input source), right click on the "Sources" box and add the appropriate input type. After attempting to add a camera, the menu in Figure 5 will pop up. This menu can also be accessed after a camera has been added by right clicking the camera and selecting "Edit". Particularly noteworthy in this menu is the ability to set the resolution and FPS (frames per second) of the camera. There are also options to flip the camera image vertically or horizontally. This is useful if a camera is most easily mounted upside down; this setting is used for the face camera.

Further customization of the camera can be performed by clicking on the "Configure" button, which brings up the menu depicted in Figure 6. All of the cameras we are using automatically adjust for the amount of light in the room and automatically focus. There appears to be no method for turning off auto-focus, but it is helpful to stop the camera from adjusting for the amount of light in the room; this is particularly helpful for the face camera. Unchecking the "Truecolor" box stops the camera from automatically adjusting the brightness, saturation, and contrast. White balance and exposure can also be manually controlled by unchecking the corresponding boxes.

We are using a date and time text overlay as part of the video capture in this experiment. This enables us to easily synchronize video data from each camera. This text overlay can be added by right clicking on the "Sources" box and selecting the appropriate option. It can be edited at a later date by right clicking on the appropriate source and selecting "Edit". When creating a new overlay or editing an existing one, the menu depicted in Figure 7 will appear.

Clicking on the "Settings" button in the OBS main menu will bring up the settings menu depicted in Figure 8. Among the configurable options are the video and audio encoding options and the save location for the files. The settings are pre-set for our experiment, so there should be little need to use this settings menu.

All of the video recordings are taken in 720p. The goal was to record in the highest possible resolution while still achieving 30 frames per second (fps). We determined that 30 fps would be enough for us to determine what the subject was doing at any point in time. Given the computing
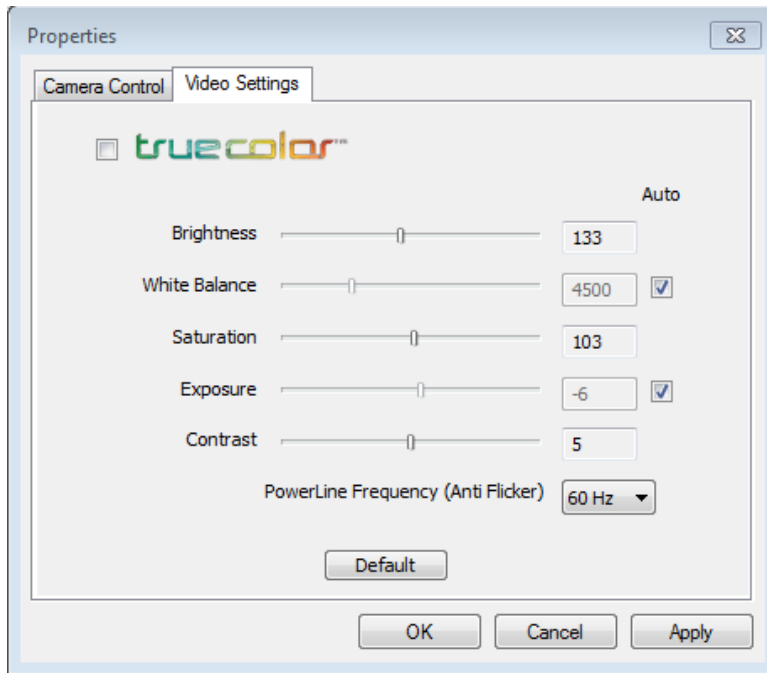
9

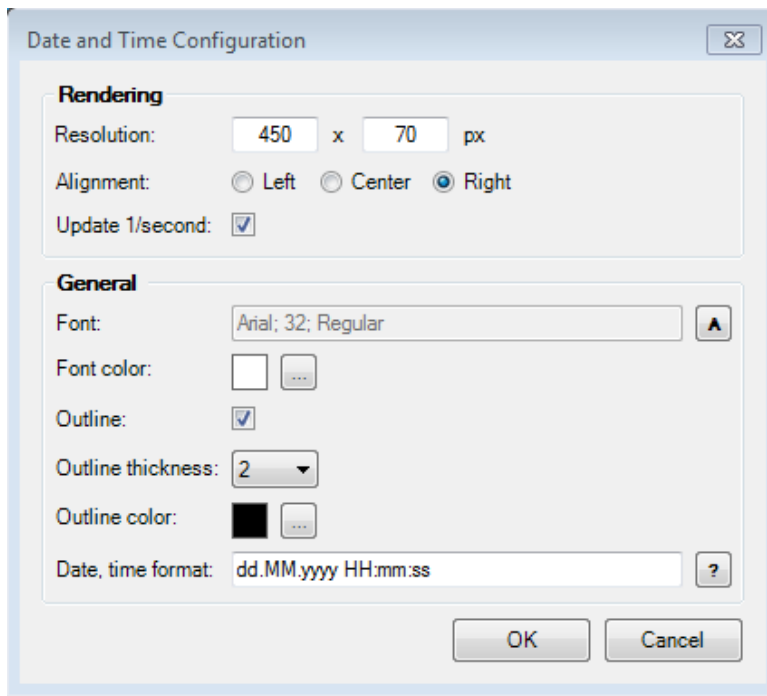Figure 6: Camera focus settings window in OBS.

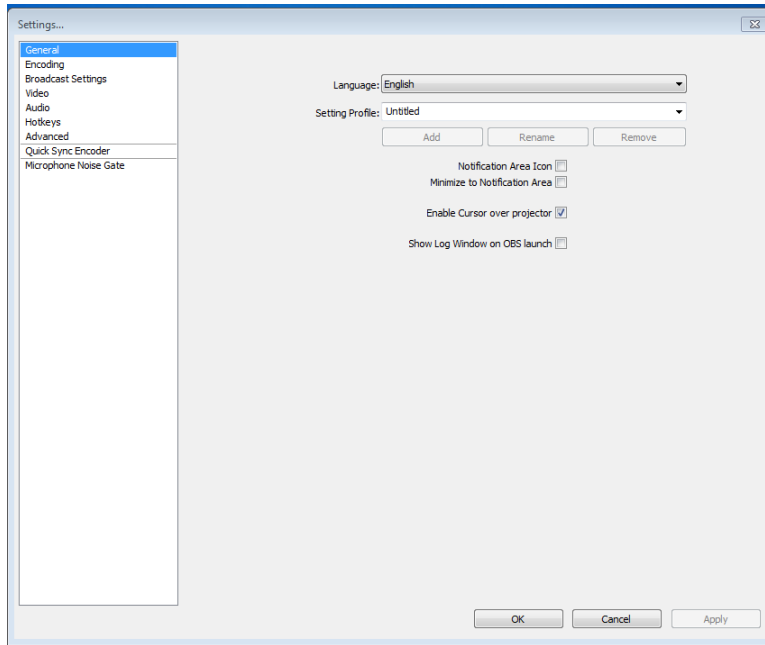Figure 7: Date and time overlay settings window in OBS.

Figure 8: Settings window in OBS.

power available to us, it would be difficult to simultaneously record 4 cameras at 30 fps and a resolution above 720p; this would likely require the use of 4 separate computers, each recording a single camera. Due to the additional complexity required by such a setup, we have opted for 720p.

OBS has several different configuration files. The default location for these files are the %APP-DATA%/Roaming/OBS folder. However, since we are using the -portable switch (see below), the configuration files are stored in the OBS installation directory. The most important configuration file is scenes.xconfig. This file contains the settings for each of the cameras, permitting fine-tuned tweaking of how each camera is recorded.

OBS also permits several different command line switches to be used. The ones that are most applicable to us are -multi, -portable, and -start. The -multi switch permits multiple instances of OBS to be run on the same machine. This is required because we are simultaneously running 4 instances of OBS, one for each camera being recorded. By design, OBS records an entire scene – consisting of one or more cameras and graphical elements – to a file. It is not possible to record individual cameras to a file unless that camera is the only item in a scene. This limitation prevents us from recording all four cameras to their own file with a single OBS instance. The -portable switch permits each instance of OBS to have its own configuration file; in our case, the configuration file is different for each instance of OBS since the camera being recorded is different as is the output file name. There are 4 OBS directories, each containing a different configuration file. Finally, the -start switch causes cameras to start recording as soon as OBS starts. This is convenient for the purposes of writing a script to start all cameras simultaneously.

The timestamp is generated using a free addon that can be found on the OBS forums.

The website for OBS is `https://obsproject.com/`. The current version of OBS being used is 0.638 Beta (released on November 11th, 2014). Support queries can be posted on the forum located at the aforementioned URL.

2.3.3 Starting and stopping cameras

Batches files have been created to start and stop the cameras. Cameras are started simply by running all instances of OBS (4 in total, one for each camera). Due to the -start switch, cameras will start recording as soon as OBS starts. Cameras are stopped by killing the OBS instances. Cameras can be started and stopped individually as well, by starting or stopping the corresponding OBS instance. It is also possible to start and stop recording without opening or closing OBS instances; this can be done by clicking the "Start recording" or "Stop recording" button on the OBS interface.

The save location for videos can be set by the user. In our case, we have set the save location to be `E:/New Videos (after 1Jan2015)`.

2.3.4 Machine specs

The video machine (named Photomon) has an Intel Xeon E5-1620v2 processor running at 3.7 GHz. The machine is equipped with 32 GB of RAM and an Nvidia Quadro K2000 video card. The current hardware in the machine was recommended to us by various members of CMU's facilities group for our specific needs. Of particular note is that an Nvidia graphics card was chosen because the OBS software permits offloading of the video encoding to Nvidia graphics cards, but not cards from other manufacturers. That said, we are actually not making use of this feature because the CPU actually has enough processing power to encode all four videos simultaneously.

In the case that additional cameras were desired or a higher resolution was desired, the best upgrades for the machine would be a faster CPU or 1-2 excellent Nvidia graphics cards. An alternative would be to purchase multiple weaker machines; each machine would then handle the load from 1-2 cameras.

We had originally considered using multiple machines, but the net cost was higher and it would be much more difficult to sync videos from multiple machines than videos from a single machine.

2.3.5 Post-processing

Post-processing of the recorded video is necessary to create the split-screen view – all four cameras in a single video file, similar to the appearance of a security camera. Post-processing of the video is done through Adobe Premiere software.

3 Physiological Measures

The physiological measures collected in the stress experiment are used for objective assessment of the affective state of the subject (i.e., neutral or stressed). Four physiological measures are being collected: (1) electrocardiogram (ECG/EKG), (2) respiration rate, (3) blood pressure, and (4) pulse rate.

3.1 Operating Procedures

Ensuring proper equipment attachment

1. Ensure that all equipment is off.
2. Ensure that the Doodad power cable is plugged in.
3. Ensure that the Critikon power cable is plugged in.
4. Ensure that the PowerLab power cable is plugged in.

5. Ensure that the pink cable is attached to the Critikon blood pressure monitor.

6. Ensure that the green cable is attached to the output port on the back of the PowerLab.

7. Ensure that the 25-pin connector is attached to the Doodad.

8. Ensure that the BioAmp is connected to the PowerLab via the I^2C port; the ports are in the back of the PowerLab and BioAmp.

9. Ensure the leads are connected to the front of the BioAmp.

10. Ensure that the Polar Respiration Belt is connected to Port 1 in the front of the PowerLab.

11. Ensure that the USB cable is plugged into the back of the PowerLab and the back of the LabChart machine.

Turning on the hardware

1. Ensure that all hardware is off.

2. Ensure that all hardware is connected.

3. Turn on the PowerLab.

4. Turn on the Doodad (by plugging in the power cable).

5. Turn on the Critikon.

6. The video and LabChart machines can be turned on at any time.

Starting up LabChart

1. Double click on `Cog_experiment` on the desktop of the LabChart machine.

2. Ensure that the preset comment miniwindow is visible.

3. Ensure that the blood pressure macros are visible (in the toolbar).

Start recording data in LabChart

1. Click one of the two start buttons (upper-right or bottom-right). Either button will work.

Stop recording data in LabChart

1. Click one of the two stop buttons (upper-right or bottom-right). The stop button replaces the start button once recording has begun. Either button will work.

Attaching the EKG cables to the subject

1. Remove the EKG cables from its storage position.

2. Ask the subject to assist you with moving their clothing out of the way.

3. Feel along the right collarbone of the subject, until you reach the point where it intersects the shoulder. Scrub the skin in this area with an alcohol wipe and attach an electrode.

4. Feel along the bottom-most rib on the left side, moving outwards, until you hit a bony intersection. Scrub the skin in this area with an alcohol wipe and attach an electrode.

5. Attach an electrode roughly parallel with the rib on the right side. Make sure the area is scrubbed with an alcohol wipe prior to electrode attachment.

6. Attach the leads by snapping them onto the electrode. The white lead goes on the right shoulder electrode, the black lead goes on the left rib electrode, and the green lead goes on the right torso electrode.

Attaching the Polar Respiration Belt to the subject

1. Ensure that you have attached the EKG electrodes and cables PRIOR to attaching the respiration belt.

2. Say to the subject: *The respiration belt wraps around your stomach, just above your belly button. The sensor* [point to the sensor]*, should be right above your belly button. The belt attaches to the sensor through a long velcro strap.* [Point out the attachment points to the subject.] *Once on, the belt should fit snugly. It should feel modestly constraining and should produce a sense of discomfort if you take a very deep breath.*

3. Ask the subject to sit down.

4. Attach the belt to the subject, with the sensor positioned just above the belly button.

Attaching the Critikon blood pressure cuff to the subject

1. Ask the subject to indicate which hand he typically operates the mouse with. The blood pressure cuff should be placed on the other arm.

2. The blood pressure cuff will be placed on the subject's upper arm, at the same level as the heart. Using a tape measure, measure the circumference of the subject's upper arm at that location.

3. Check if the currently attached cuff is the right size. If it is not, swap it out for an appropriately-sized one.

4. If the subject is wearing any clothing that is not skin-tight, ask them to move the fabric out of the way so that the cuff can be attached (e.g., by rolling up a sleeve).

5. Ensure that the hose connected to the cuff is not kinked or warped in any way.

6. Place the cuff so that the subject's artery is aligned with the cuff arrow marked "artery."

7. Squeeze the cuff to remove all air from it.

8. Wrap the cuff snugly around the subject's limb. Ensure that the cuff index line falls within the range markings. If it does not, use a larger cuff and repeat the attachment process from the beginning. Ensure that the velcro attachments are secure. You should be able to fit a single finger between the subject's arm and the cuff.

Ensuring the EKG is functional

1. Check to see that the characteristic QRS wave is present and has little noise.

2. If no QRS wave is present, check that all connections (leads to electrodes, leads to BioAmp, and BioAmp to PowerLab) are firmly connected.

3. If no QRS wave is present and all the connections are firm, ensure that the range is set to 100 mV.

4. If the QRS wave is inverted, check that the black and white leads are clipped to the proper electrodes. The white lead should be clipped on to the right shoulder electrode, and the black lead should be clipped on to the left rib electrode. If the QRS wave is still inverted, switch the black and white leads and make a note of it. (This can occur due to an unusual medical condition that causes the heart to be rotated. It should occur in about 2-3 people per 100.)

Ensuring the Polar Respiration Belt is functional

1. Ask the subject to deeply inhale and then exhale. This should produce a clear and obvious signal.

2. If no signal is present, check that the connection between the respiration belt and the Pow-erLab is firm.

3. If there is still no signal present, ensure that the range is set to 50 mV.

Ensuring the Critikon is functional

1. Ensure that there are no warnings displayed. If there are, restart the Critikon to clear the errors.

2. Take a blood pressure reading and ensure that a proper reading is displayed. The systolic blood pressure should be less than 140 and the diastolic blood pressure should be less than 90, unless the subject has high blood pressure. The pulse rate should be between 60 and 100 for most subjects. This may be lower if the subject exercises heavily.

Removing the EKG from the subject

1. Ask the subject to remove the electrodes, with leads still attached, from their body.

2. Unclip the leads from the electrodes.

3. Provide the subject with alcohol wipes if they wish to clean off any residue.

Removing the Polar Respiration Belt from the subject

1. Remove the EKG leads and electrodes prior to removing the belt.

2. Detach the velcro strap and remove the belt.

3. Clean the belt and curl it up prior to storage.

Removing the Critikon blood pressure cuff from the subject

1. Ensure the the cuff is not currently inflated.

2. Detach the velcro and remove the cuff. Ensure that the hose is not kinked when the cuff is put into storage.

3.2 Troubleshooting

For all problems, note the problem in the checklist and inform Shing-hon and Roy asap.

A physiological measure is not reading a signal

	When this happens	What to do
1.	There is no discernible reading	Narrow the range by clicking the arrow next to the channel reading and selecting: **100mV for EKG, 50mV for Respiration belt**
2.	There is no discernible reading even though the range has been reduced	Double-check to see if the EKG leads and the respiration belt are attached securely.
3.	There is no discernible reading even though the range is reduced and the equipment is secure. Also, there are five minutes remaining in the rest period or during the Purple task.	Save the file and restart LabChart
4.	There is no discernible reading even though LabChart was restarted	Reboot LabMon[1] and restart LabChart.
5.	There is no discernible reading even though LabMon has been rebooted	Continue with the experiment with the working physiological measures

---

[1]login: administrator, password: adm 020295

A physiological measure is out of bounds or has an odd reading

	When this happens	What to do
1.	Reading is out of bounds	Adjust the range by clicking the arrow next to the channel reading and selecting a wider range: **100mV for EKG, 50mV for Respiration belt**
2.	Odd reading but not out of bounds	Insert a comment in LabChart
3.	Odd reading before the **1st rest period**	Re-attach equipment
4.	Odd reading while the subject is **not typing**	Ask subject to check all electrodes are attached and the respiration belt is snug on their chest – if this doesn't work, continue with the experiment anyway
5.	Odd reading while the subject **is typing** or using Purple	Wait until the subject completes the typing task or Purple. Then, ask the subject to check all electrodes are attached and the respiration belt is snug on their chest – if this doesn't work, continue with the experiment anyway.
6.	Odd reading is fixed	Insert comment into LabChart saying the odd reading was fixed.

An EKG electrode has fallen off

	When this happens	What to do
1.	EKG electrode fell off	Insert a comment in LabChart
2.	EKG electrode fell off **before the 1st rest period**	Attach a new electrode to the subject, then the EKG lead.
3.	EKG electrode fell off **when the subject is not typing or using Purple**	Attach a new electrode to the subject, then the EKG lead.
4.	EKG electrode fell off **while the subject is typing or using Purple**	Wait until the next rest period and then attach a new electrode to the subject, then the EKG lead.
5.	After the EKG electrode is replaced	Insert a comment in LabChart

Subject complains that the respiration belt is too loose/tight

	When this happens	What to do
1.	Subject complains that the respiration belt is too loose/tight	Insert a comment in LabChart
2.	Subject complains **before the first rest period**	Ensure the subject is seated and re-adjust the respiration belt
3.	Subject complains **when the subject is not typing or using Purple**	Ensure the subject is seated and re-adjust the respiration belt
4.	Subject complains **when the subject is typing or using Purple**	Wait until the next rest period, then, ensure the subject is seated and re-adjust the respiration belt
5.	After respiration belt adjustment is made	Insert a comment in LabChart

Respiration belt falls off

	When this happens	What to do
1.	Respiration belt falls off	Insert a comment in LabChart
2.	Respiration belt falls off **before the first rest period**	Ensure the subject is seated and re-attach the respiration belt
3.	Respiration belt falls off **when the subject is not typing or using Purple**	Ensure the subject is seated and re-attach the respiration belt
4.	Respiration belt falls off **when the subject is typing or using Purple**	Wait until the next rest period, then, ensure the subject is seated and re-attach the respiration belt
5.	After respiration belt is re-attached	Insert a comment in LabChart

BP cuff is too loose/tight

	When this happens	What to do
1.	Subject complains that the BP cuff is too loose/tight	Check the cuff size and re-adjust or swap with an appropriate-sized one.
2.	Subject complains **before the first rest period**	Adjust the BP cuff.
3.	Subject complains **when the subject is not typing or using Purple and no BP reading will be taken in the next minute**	Adjust the BP cuff.
4.	Subject complains **when the subject is not typing or using Purple BUT a BP reading will be taken in the next minute**	Wait until after the BP reading to adjust the cuff.
5.	Subject complains **when the subject is typing or using Purple**	Wait until the next rest period, then adjust the cuff.

BP cuff has fallen off

	When this happens	What to do
1.	The BP cuff falls off subject's arm	Check the cuff size and re-adjust or swap with an appropriate-sized one.
2.	The BP cuff falls off **during a rest period**	Adjust the BP cuff immediately and take any missed BP readings immediately after adjustment.
3.	The BP cuff falls off **in a non-rest period**	Re-attach it before the next rest period.

Critikon alarm has sounded

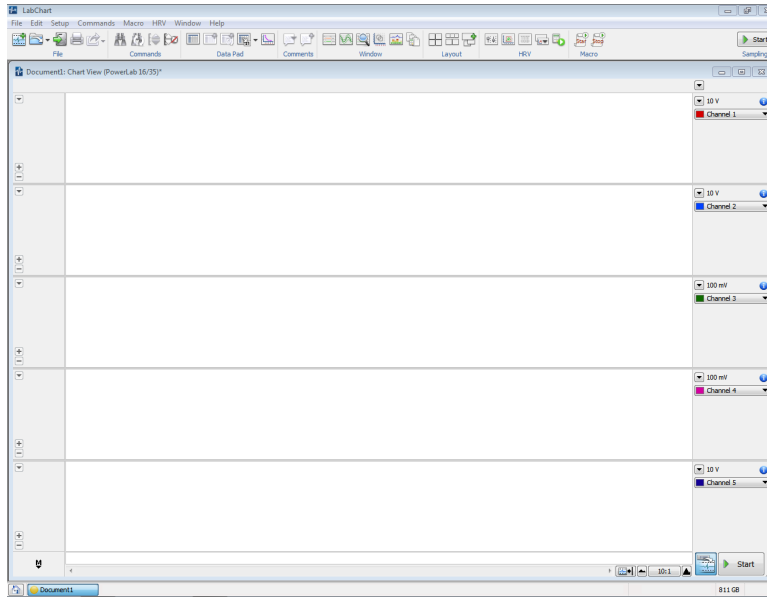|	When this happens	What to do
1.	Critikon alarm has sounded	Press the alarm button twice; the first press silences the alarm temporarily and the second silences it completely.
		Ensure the subject is ok. The alarm only sounds when the BP is excessively high or low. Call x82323 (Campus EMS) if necessary.
		If the subject is ok, discontinue the use of the BP monitor and complete the experiment.

Figure 9: LabChart's main window.

Critikon printer has jammed

	When this happens	What to do
1.	Critikon printer has jammed	Open the printer and remove the jam.
		Ensure that a small strip of paper is sticking out of the printer to prevent future jams.

3.3   Technical Details

ECG/EKG is collected through a 5-lead electrode system, which feeds data to the PowerLab. Respiration rate is collected using the Polar Respiration belt, which also feeds data to the PowerLab. Blood pressure and heart rate are collected by the Critikon. The output of the Critikon is directly printed out using a printer attached to the device. However, each reading is trigged by LabChart, via the Doodad. LabChart is the name of the software that interfaces with the PowerLab hardware.

The LabChart software is responsible for integrating all of the physiological signals in a time-synchronized fashion, and then recording it into a single file. The main screen for LabChart is displayed in Figure 9. At the top left of the main window is the menu bar; the contents of each sub-menu in this bar is described later. The center of the screen contains channels of input. In this example, there are 5 channels in total. In the right sidebar for each channel, the channel range is displayed. This range can be altered by clicking on the drop-down arrow next to the range. In this example, Channels 1, 2, and 5 have a range of 10V, while Channels 3 and 4 have a range of 100 mV. Other options can also be adjusted in this sidebar; however, we will not be altering these options for this experiment. Data collection can be started and stopped by clicking on one of the two start/stop buttons. One is located in the upper-right, while the other is located in the bottom-right. When data collection is active, the data will scroll across each of the channels.

The setup sub-menu, depicted in Figure 10, allows many settings to be altered. For our experiment, we will utilize Channel Settings, Digital Output, and Preset Comments. Clicking on Channel Settings will bring up the window shown in Figure 11. From here, it is possible to set
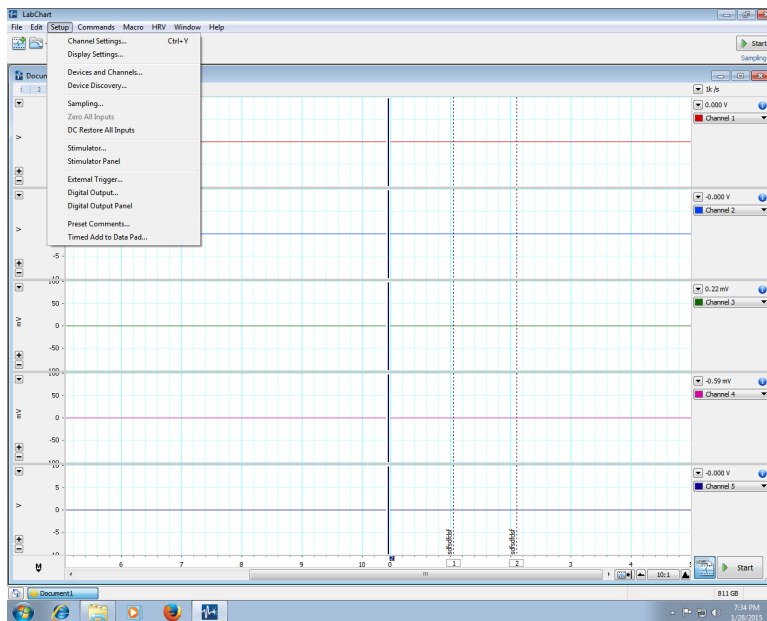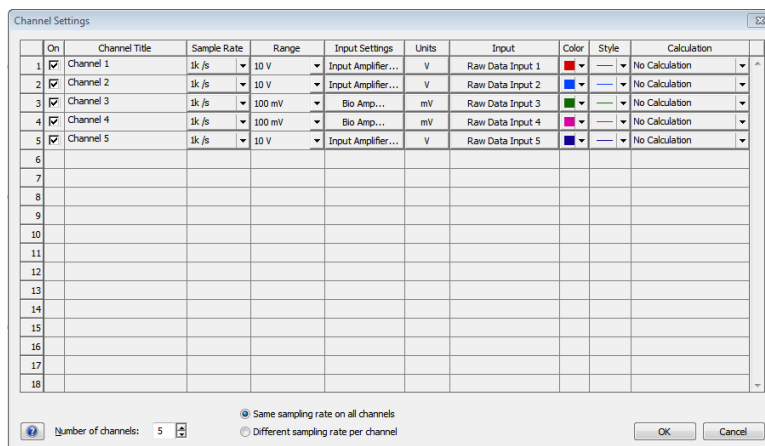
Figure 10: LabChart's setup menu.
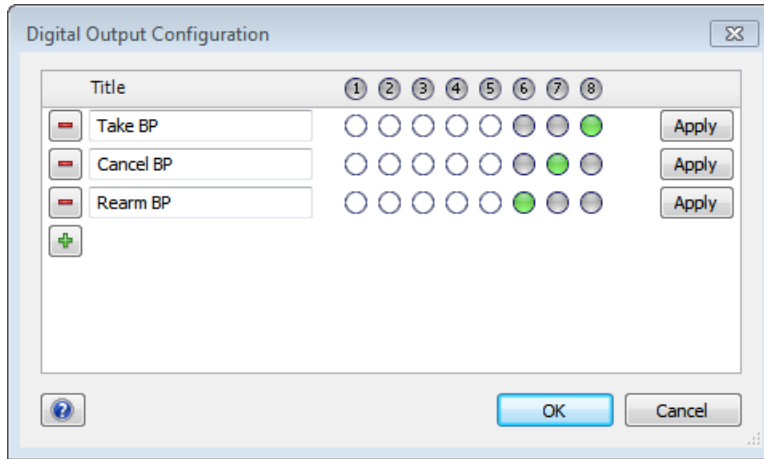

Figure 11: LabChart's channel settings menu.

Figure 12: LabChart's digital output menu.

many parameters regarding the channels. For example, the number of channels, their input and/or associated calculation, the range, the sample rate, and the units, can all be set via this window.

The digital output window, depicted in Figure 12, allows us to control the signal emitted from the digital output port in the back of the PowerLab. For this experiment, we initially intended for there to be one-way communication with the Critikon, via the Doodad. However, we later discovered that this was not worth the effort. Nevertheless, here are Harvey's instructions regarding communication with the Critikon:

```
Pin 1:  initiate BP start , if the trigger (see pin 3) is armed.
Nothing otherwise.
Pin 2:  cancel an in-progress BP sequence. Nothing if a BP sequence
isn't running.
Pin 3:  trigger arm/print results:  After starting a BP sequence and
either cancelling or finishing it normally, assert this pin to re-arm
the "start" pin (pin 1). This acts as a "debounce"/multiple-start
safety to avoid sending redundant start commands to the BP machine.
It will also cause a PRINT command to be sent to the BP machine to
print out the latest result
```

Note that Harvey's numbering of the pins is mirrored to the LabChart numbering. That is, what he calls Pin 1/2/3 is labeled as Pin 8/7/6 in LabChart. A pin is considered "on" if it exceeds 2.5V; this is achieved by activating the corresponding digital output pin. Similarly, a pin is considered "off" if it falls below 2.5V.

We do not directly control the digital output signal, since the timing on rearming is quite sensitive. Keeping Pin 3 (as per Harvey's instructions) on for more than a few moments will cause multiple printings of the last blood pressure reading. Instead of controlling the digital output directly, we instead employ LabChart's macro functionality. Two macros have been created. One rearms the Critikon and takes a blood pressure reading; this has the side-effect of causing the last reading to be printed. The second cancels the current reading. The macros also insert comments into the LabChart file to indicate when a reading is started or stopped. The macro window can be accessed via the Macro sub-menu (see Figure 13 and selecting "Manage...". The macros can be viewed in Figures 14 and 15.
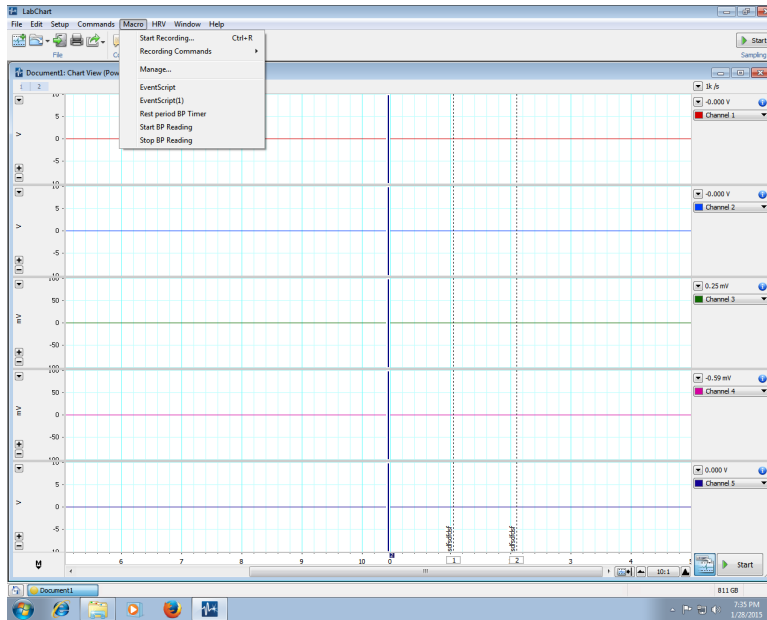
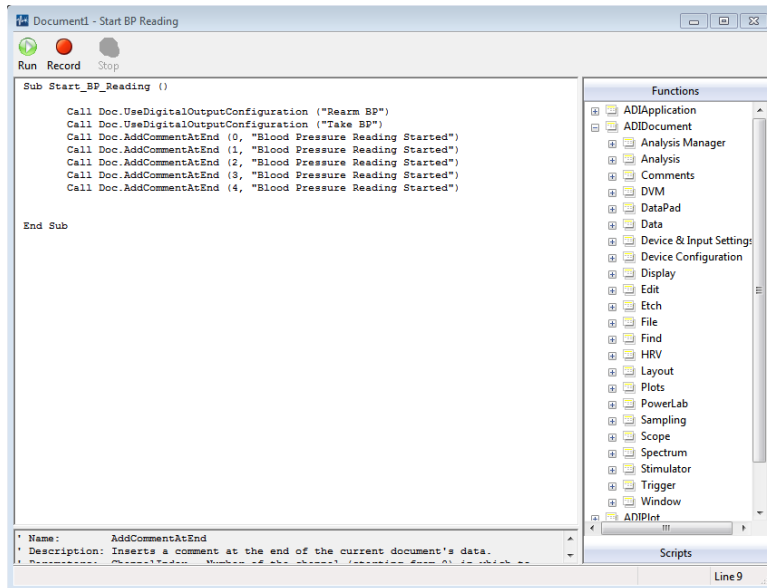Figure 13: LabChart's macro sub-menu.

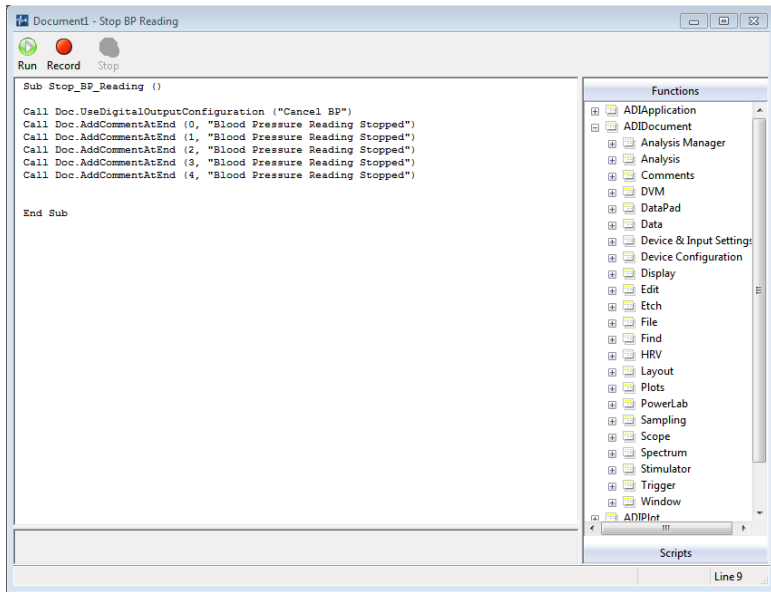Figure 14: LabChart macro to start a blood pressure reading.

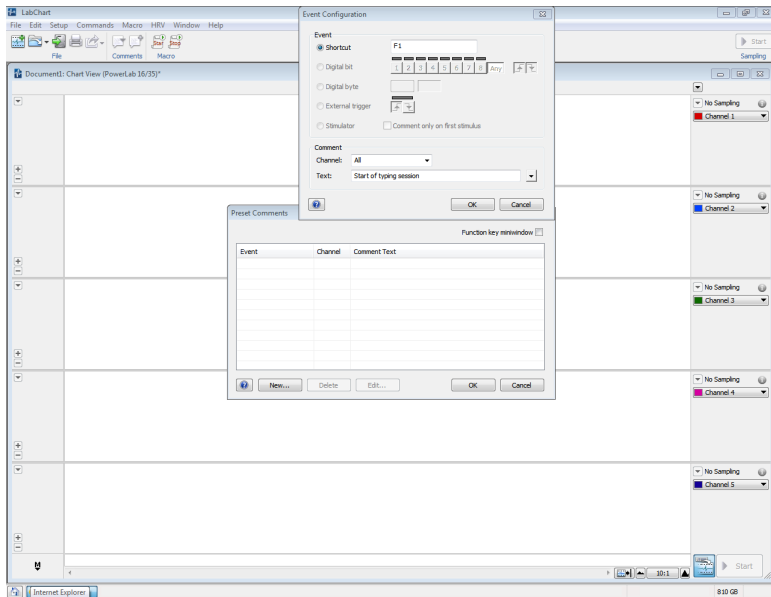Figure 15: LabChart macro to stop a blood pressure reading.

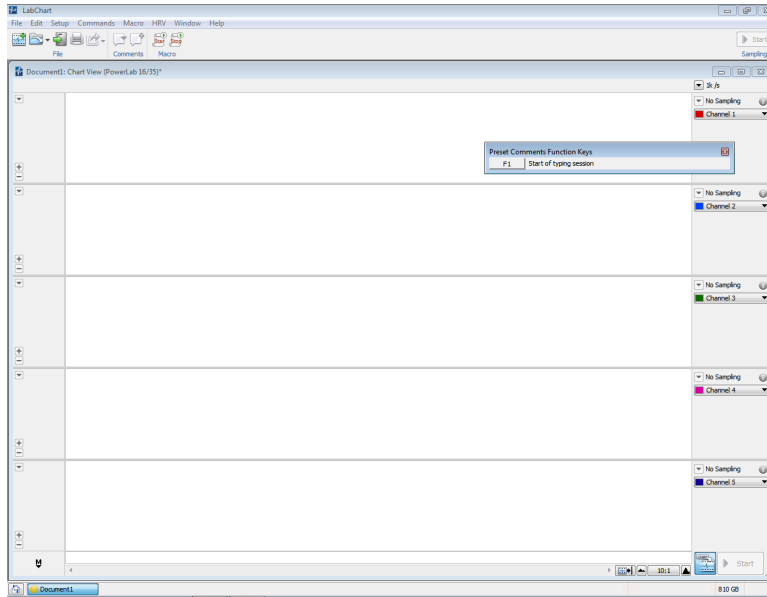Figure 16: LabChart's preset comments window.

Figure 17: LabChart's preset comments miniwindow.

The final item of interest in the setup menu is the Preset Comments window, depicted in Figure 16. As the name suggests, this menu allows the user to define comment strings that can then be inserted into the LabChart file with the press of a single button. This allows us to record when a subject begins or ends a typing session or rest period. Once preset comments have been defined, they can be inserted into the file by either clicking on the comment in the miniwindow (see Figure 17) or by hitting the appropriate hotkey. In Figure 17, the F1 hotkey has been assigned to the inserting the comment "Start of typing session".

3.3.1 Physiological Equipment

The first decision made with regard to the physiological equipment used in this experiment was to choose the PowerLab hardware and corresponding LabChart software. This combination is manufactured by AD Instruments. This choice was made largely because the equipment could perform the necessary physiological monitoring and because it had a non-proprietary data format. Other equipment that we considered had proprietary data formats, making it difficult to share data with other researchers and making it difficult to perform our own analysis.

There are four physiological signals that are recorded as part of this experiment: 1) electrocardiogram (ECG/EKG), (2) respiration rate, (3) blood pressure, and (4) pulse rate. All signals are routed through the PowerLab and then into a computer running LabChart. The EKG is collected using a 5-lead system (of which we use 3 leads), which is routed through a BioAmp (also manufactured by AD Instruments); the BioAmp is connected to the PowerLab. Respiration rate is collected using a Polar respiration belt kit, which plugs directly into the PowerLab. Blood pressure and pulse rate are both collected by a Critikon blood pressure monitor.

The EKG and respiration rate data are automatically synchronized by LabChart. It was our original intent to also synchronize the data captured by the Critikon monitor. We asked Harvey to manufacture a device (dubbed the Doodad) to facilitate this. The idea was to have LabChart automatically trigger a blood pressure and pulse reading at fixed times throughout the experiment. The resulting data from this reading would then be sent back to LabChart, where it would be au-

tomatically recorded. The automation was designed to reduce the burden on the experimenter and hopefully reduce the potential for mistakes. Due to inconsistencies with the Critikon programming, the Doodad is unable to send back the data. However, the automated triggering still works.

4 Typing data

As in prior keystroke experiments, typing data will be collected using MetriTextPrompter (MTP). The keyboard used will be an Apple keyboard connected to the Gizmo. Subjects will type the same phrase repeatedly for each sessions.

4.1 Operating Procedures

Starting the warm-up session

1. Bring up the command window on the laptop.
2. Run `run-cog-warmup.bat`.

Starting the first neutral session

1. Bring up the command window on the laptop.
2. Run `run-cog-n1.bat`.

Starting the cognitive load session

1. Bring up the command window on the laptop.
2. Run `run-cog-cog.bat`.

Starting the second neutral session

1. Bring up the command window on the laptop.
2. Run `run-cog-n2.bat`.

4.2 Troubleshooting

The participant is distracted mid-repetition.

When this happens	What to do
1. The participant stops typing mid-repetition	Make a typo so that the data entry is invalid.
	If it's too late to make a typo, record the counter number in the notes section of that typing session.

4.3 Technical Details

Technical details have thoroughly detailed in other operations manuals. See the Strong, Hester, or Phone operations manual for details.

5 Neutral induction

Neutral induction is performed twice during the course of the stress experiment. Its purpose is to establish that the subject is actually in a neutral state; the validity of the experiment would be compromised if we simply assumed the subject was in a neutral state. The first induction is done prior to the first typing baseline sample, to ensure that the subject is in a neutral state. The second induction is performed after the stressed typing and before the post-stress baseline sample.

5.1 Operating Procedures

1. Give the subject instructions for the neutral induction
2. Turn on the relaxing video
3. Provide subject instructions for the simple task – count the number of types of animals for the first video; count the number of human divers for the second video
4. Provide the subject with pen and paper to perform the task
5. Set a timer for 30 minutes
6. During the 30 minute period, check the physiological measures at least once a minute to ensure they are still functioning
7. At the 21 minute mark, take a blood pressure reading
8. At the 24 minute mark, take another blood pressure reading
9. At the 27 minute mark, take another blood pressure reading
10. At the 30 minute mark, take a final blood pressure reading

5.2 Troubleshooting

For all problems, note the problem in the checklist and inform Shing-hon and Roy asap.

Some distraction has occurred

	When this happens	What to do
1.	A distraction occurred that **did not require the subject to leave the room** (e.g. People talking loudly in the hallway)	Attempt to stop the distraction asap.
		Make a note of the time and nature of the distraction in checklist
		Continue with the experiment.
2.	A distraction occurred during the rest period that **required the subject to leave the room for < 5 min** (e.g. bathroom break)	Make a note of the time and nature of the distraction.
		Resume the rest period and add 10 minutes to the end.
		Continue with the experiment
3.	A distraction occurred during the rest period that **required the subject to leave the room for > 5 min** (e.g. fire alarm)	Restart the rest period.
		Continue with the experiment

A blood pressure reading was missed

	When this happens	What to do
1.	Missed a BP reading and the next scheduled BP reading is **more than 1 min**	Take the BP reading asap and make a note in checklist.
2.	Missed a BP reading and the next scheduled BP reading is **less than 1 min**	Skip the BP reading and note in checklist that the BP reading was missed.
3.	Missed two BP readings	Take the BP reading asap and note in checklist that the two BP readings were missed.

A form was not administered

	When this happens	What to do
1.	A form was omitted from the experiment	Note omission in checklist and notify Roy/Shing-hon.

6   Stress induction

Stress induction is performed once during the course of the experiment. It is done using a combination of the Purple framework and social evaluation.

6.1   Operating Procedures

1. Give the subject instructions for using the Purple software
2. Start up the Purple software
3. Start a timer for 15 minutes
4. At the 5 minute mark, perform social evaluation
5. At the 10 minute mark, perform social evaluation
6. Take a blood pressure reading

6.2   Troubleshooting

For all problems, note the problem in the checklist and inform Shing-hon and Roy asap.

The Purple software stopped working

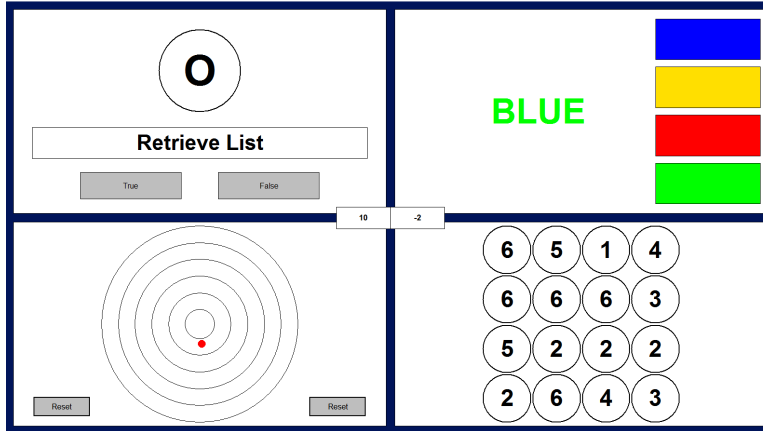|	When this happens	What to do
1.	The Purple software crashes	Close and re-start Purple.
		Inform subject that we need to re-start the Purple session.
		Make a note in the checklist.
2.	The Purple software crashes twice	Skip the Purple session.
		Continue with the remainder of the experiment, starting with the typing session

Figure 18: Purple's main window.

6.3 Technical Details

The Purple software was originally provided to us by Mark Wetherell. The software is intended to be a pseudo-realistic simulation of a typical work environment, where an individual is being asked to multi-task heavily. Studies by Mark and his colleagues have demonstrated that using the Purple framework induces a mild to moderate stressor in subjects. In our experiment, we are combining the use of the Purple software with social evaluation, in the form of negative feedback from the experimenter. We have a re-written version of the Purple software to increase the stress on the user. For example, some bugs that caused time-outs to not happen have been fixed. The penalty for time-outs has also been increased to force subjects to attend equally to all the tasks; in pilot studies, we found that subjects tended to ignore some modules since the time-out penalties were not severe.

The Purple framework consists of 4 different quadrants, each with a task that the subject must perform. The difficulty level (speed) can be varied, as can the tasks themselves. Figure 18 shows the main window for Purple. The upper-left corner contains the task intensity selection menu. In our experiment, we will use low for the familiarization task and high for the actual task. This will allow the subject to have an easy time learning how to use the software during the familiarization period, while making the actual task difficult enough to induce cognitive load. The bottom-left corner contains the task duration menu. In our experiment, we will use 2-minutes for the familiarization period and 15 minutes for the actual task. In the top-center of the screen is the results file location. The middle-center of the screen contains the toggling for "vs mode", where a score appears on the screen that is purported to be from a human opponent, but is in actually controlled by the software. The bottom-middle of the screen contains options for the scoring; we allow the subject to see the score and allow negative scores. The right-hand side of the screen contains saving and loading configurations and the selection of modules (see Figure 19). In our experiment, we have chosen Numbertap, Stroop, Tracking, and LetterSearch as the four tasks. These tasks are chosen for consistency with what Mark has used in his studies.

7 Forms

There are several different types of forms will be employed during the course of the experiment: Consent form, Demographic survey, Perceived Stress Scale, Long-form STAI (Y-2), State-Anxiety
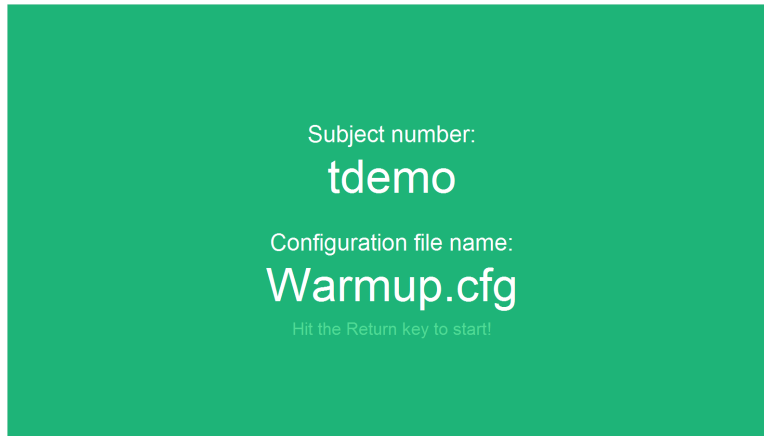
Figure 19: Purple's configuration window.

Visual Analogue Scale, and the NASA-TLX workload inventory.

7.1 Operating Procedures

Consent form

1. Give a copy of the consent form to the subject.

2. *The first step in this experiment is to obtain your informed consent. I will describe the key points to you now, and then I will give you an opportunity to read through the form itself.*

3. *You must be at least 18 years old to participate in this experiment. You must be fluent in English, have at least three years of experience typing on a computer, and must type at least 30 words a minute.*

4. Ask the subject to confirm that all four facts are true.

5. *You must not have any history of cardiac or neurological disorders, anxiety or stress disorders, or sleep disorders.*

6. Ask the subject to confirm they do not have any such history.

7. *You must not have a systolic blood pressure above 140 or a diastolic blood pressure above 90. You also must not have had a stroke. We will be verifying your blood pressure in a moment, if you are unsure about your blood pressure readings.*

8. Ask the subject to confirm that they have not had a stroke and that they believe their blood pressure lies in the appropriate range.

9. *It is also important that you have not had excessive caffeine consumption (more than 3 cups of coffee in a day) and psychoactive drugs for a 72-hour period before the experiment.*

10. Ask the subject to confirm that they have not consumed excessive caffeine or psychoactive drugs for the 72-hour period.

11. *The risks and discomfort associated with participating in this study are no greater than that you would encounter in your daily life or during typical office work.*

12. *There will be no personal benefit to you for participating in this study. However, we will provide compensation for your time at the completion of this study. The compensation will be 60 dollars, in cash. You will earn 10 dollars for completing the experiment, and an additional*

bonus of 50 dollars if you are highly engaged during the study. If you choose to leave early, you will only the 10 dollars.

13. *During the course of this study, we will be collecting a variety of data from you. This includes your written responses to questionnaires and forms, physiological measurements, and audio-visual data. All data collected from you will be protected by storing it in a locked location on Carnegie Mellon property and will not be disclosed to third parties. Personal identifiers, like your name, phone number, and date of birth, will be kept separate from other information collected.*

14. *Your participation in this study is completely voluntary. You are free to leave at any time.*

15. *Please take a moment to read through the form now. Take as long as you would like. Please feel free to ask any questions you may have. Please initial where called for in the form and then sign your name at the end of the form.*

16. Give the subject as long as they wish to read through the consent form.

17. Answer any questions the subject might have.

18. Ensure that the subject initials where called for in the form.

19. Ensure that the subject signs and dates the consent form under "Participant Signature".

20. Sign and date the form under "Signature of Person Obtaining Consent".

21. Place the consent form into the manila folder.

Demographic survey

1. Give a copy of the demographic survey to the subject.

2. *The next form is a demographic survey. This asks questions like: "Are you right-handed or left-handed?" and also questions about your typing habits.*

3. *Please take some time to fill out this form now. Please let me know if you have any questions.*

4. Place the demographic survey into the manila folder.

Perceived Stress Scale

1. Give a copy of the Perceived Stress Scale to the subject.

2. *The last form we have for now is the Perceived Stress Scale. For each question, please circle a number to indicate how often you felt or thought a certain way in the past month.*

3. Place the Perceived Stress Scale into the manila folder.

Long-form STAI Y-2

1. Give a copy of the Long-form STAI to the subject.

2. *The next form is the Long-form State Trait Anxiety Inventory, or STAI. Please circle the number corresponding to statement describing how you feel right now, at this moment.*

3. Allow the subject to respond to all the questions.

4. Ensure that the subject has answered all of the questions.

5. Place the long-form STAI Y-2 into the manila folder.

State-Anxiety Visual Analogue Scale

1. Give a copy of the VAS to the subject.

2. *I will ask you to fill out this form on a few occasions throughout the experiment. Please make a mark on each line indicating how you feel at this moment. Please look over it now and see if you have any questions.*

3. Answer any questions the subject might have.

NASA-TLX Workload Inventory

1. Present a copy of the NASA-TLX to the subject.

2. *Throughout the experiment, I will ask you to fill out this form on a few occasions. For each question on the form, place a vertical mark on each line corresponding to how you feel at the moment.*

7.2 Troubleshooting

For all problems, note the problem in the checklist and inform Shing-hon and Roy asap.

Subject wants to correct an answer on the form

	When this happens	What to do
1.	Subject wants to correct an answer on the form **while completing the form**	Ask subject to mark the incorrect answer with a large **X**.
2.	Subject wants to correct an answer on the form **after completing the form**	Tell subject that they cannot change answers once they have completed the form.

7.3 Technical Details

The consent form is mandatory due to IRB regulations. The demographic form is used to obtain demographic information from our subjects; this information could have explanatory capabilities for changes in a subject's typing. The remaining forms (long-form STAI, PSS, short-form STAI, and Bond-Lader VAS) are chosen because they are the standard forms in stress research. The long-form STAI and the PSS covers the way that stress impacts the subject in terms of their general activities. The short-form STAI and Bond-Lader VAS cover short-term responses to stress, measuring changes in stress due to the experimental conditions. These forms were also used by Mark Wetherell in his own research.

8 Still camera

We will be using a remotely-activated still camera to take KPECS pictures for this experiment.

8.1 Operating procedures

Turning on the still camera

1. Turn on the camera.

2. Turn on the remote attached the camera.

3. Detach the remote from the camera itself.

Turning off the still camera

1. Attach the remote to the camera.

2. Turn off the remote.

3. Turn off the camera.

8.2   Troubleshooting

The camera or remote does not function.

When this happens	What to do
1. The camera or remote will not turn on.	Replace the batteries for the device that will not turn on.
	If the scale still does not turn on, skip the still photos and let Roy and Shing-hon know about the issue.
2. The remote does not work.	Ensure that the remote is turned on.
	Replace the batteries in the remote if it does not turn on.
	If the remote still does not work after replacing the batteries, take the pictures manually by pressing the button on the camera. Inform Roy and Shing-hon about the issue after the experiment is over.

8.3   Technical Details

9   Height and weight measurements

We will be measuring subjects' height and weight to help calibrate the EKG readings.

9.1   Operating procedures

Measuring the subject's height

1. Ask the subject to remove their shoes.

2. Ask the subject to stand against the door with their heels against the door.

3. Ask the subject to place a hand on the top of their head against the scalp.

4. Record the reading on the measuring tape. The offset from the scale starting off the floor will be added in later.

Measuring the subject's weight

1. Turn on the scale by pressing the button on the front.

2. Ask the subject to remove their shoes.

3. Ask the subject to step on the scale and to avoid holding onto or leaning on anything.

9.2   Troubleshooting

The subject is too tall.

	When this happens	What to do
1.	The subject is taller than the top of the scale.	Estimate the subject's height to the best of your ability.
		Make a note that the measurement was an estimate due to the subject being taller than the top of the scale.

The scale will not turn on.

	When this happens	What to do
1.	The scale will not turn on.	Replace the batteries for the scale.
		If the scale still does not turn on, skip the weight measurement and let Roy and Shing-hon know about the issue.

9.3 Technical Details

The scale is designed to automatically turn off after a period of inactivity. There is no need to manually turn off the scale.

The height chart is offset from the ground by approximately 18 inches. The reading should be taken according to what is on the chart; adjustments to the height to accommodate the offset will be done at a later time.