Evaluating differences between keystroke datasets collected under lab and field conditions

Wangzi He

CMU-CS-15-129
August 2015

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Roy Maxion, Chair
Daniel Siewiorek

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Keywords: Keystroke dynamics, Keystroke biometrics, Computers, Science

Abstract

Background. Keystroke dynamics is the process of measuring and assessing a user's typing rhythms based on durations of single-key presses and key-to-key transitions. Anomaly detectors are frequently employed to distinguish between legitimate users and impostors based on differences in their typing rhythms. Previous research has used both field and lab data to evaluate these detectors, but no study has yet established the difference between lab and field data, nor the impact of such differences on evaluation results.

Aim. We compare, on various dimensions, two data sets that are identical in all respects except one – one data set was collected under laboratory conditions, and the other under field conditions. We develop a methodological framework that can be used to compare field and lab versions of datasets.

Data. Each data set, lab and field, consisted of 51 subjects who typed 400 repetitions of the password ".tie5Roanl" over 8 sessions of 50 repetitions each.

Methods. Distributional differences in demographic data and typing-time data were assessed using statistical tests and regression analysis.

Results. Findings include: (1) field and lab populations were similar in terms of demographics, although the field population contained more subjects in their early 20s; (2) lab and field data were statistically different in terms of total time to type the password, but the effect size was small; (3) field subjects generally have faster rates of typing-skill acquisition; (4) a set of top-performing anomaly detectors produced nearly the same results when evaluated on lab and field data.

Conclusions. While some statistically significant differences may exist between lab and field data, these differences have minimal impact on evaluations of anomaly detectors. Results suggest that detectors developed using lab data are likely to have equivalent performance when deployed in the field.

Acknowledgments

I could not have completed this thesis without the help of many colleagues, family, and friends. First and foremost, I want to give my heartfelt thanks to my advisor, Roy Maxion. Roy has been an inspirational role model; his passion for research and aim for perfection will always inspire me. I am grateful for everything I learned from him. I am also privileged to have had Dan Siewiorek on my thesis committee, and wish to thank Dan for his thoughtful comments during the review of this thesis. I feel extremely fortunate to have worked with the best researchers in this field. Shing-hon Lau provided help and valuable suggestions since the very beginning of my work. I deeply appreciate the time Shing-hon spent on helping me improve this thesis. Pat Loring was instrumental, patiently teaching me details of running experiments.

My mother and father – Tong Wang and Jing He – have always encouraged me. Their support powered me through the long journey abroad. Many friends also provided help throughout my time at CMU: William Park, Arnob Mallick, Jaeah Kim, Shashank Singh, Jason Harding, Aashish Jindia, Andrew Audibert, Leo Scholl, Mike Rosen, Shikun Zhang, Shuaihang Wang and Roy Zou. As should be evident from the long list of names, there are many who supported me in this work. Without their help, I could not have done it. I am honored to have such great friends, family, and colleagues.

谨以此论文献给亲爱的爸妈。感谢爸妈多年来对我的辛勤栽培！

Contents

Chapter 1

Introduction

Keystroke dynamics, also known as keystroke biometrics, is the term given to the procedure of measuring and assessing a user's typing rhythm. These typing rhythms are composed of hold times – the duration a single key is depressed – and latency times – the duration between consecutive key presses – as depicted in Figure 1.1. Anomaly detectors are often employed to distinguish between a single legitimate user and multiple impostors on the basis of individual differences in these hold and latency times. Evaluating the abilities of different anomaly detectors to accomplish this task has been the focus of considerable research efforts.

Evaluating the performance of a set of detectors requires typing data, and researchers have collected data under both lab and field conditions. Lab data is generally viewed as being of higher quality, as it is collected under tightly-controlled conditions. However, collection of lab data limits the pool of potential subjects to a small geographical radius around the laboratory and requires the presence of a research assistant to facilitate data collection, restricting the hours where data can be collected and greatly limiting the number of subjects that can be run in a given time frame. The collection of field data is not bound by these limitations. The pool of potential subjects includes anyone with an Internet-connected computer; data can be collected at any time of day and multiple subjects can provide data simultaneously. Unfortunately, no research has yet compared the quality of field and lab

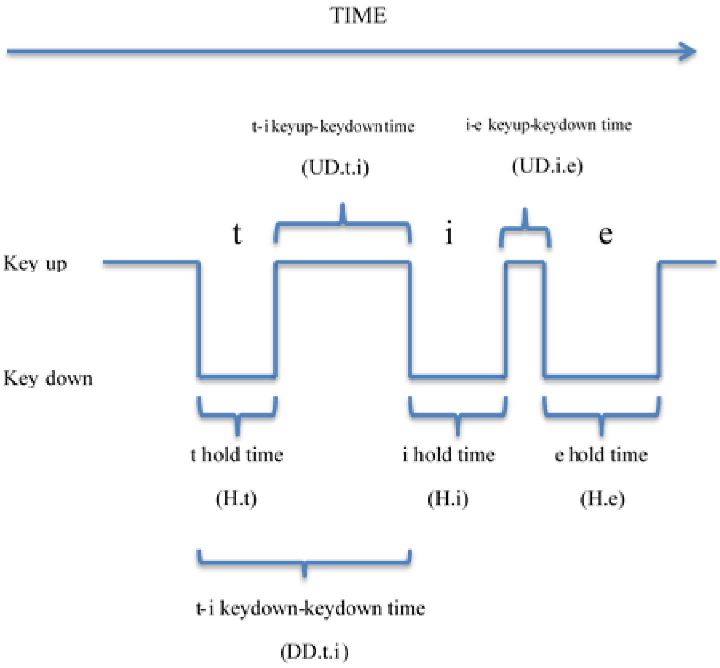


Figure 1.1: Feature descriptions.

data; therefore, it is uncertain what degradation of evaluation accuracy may result from using field data instead of lab data.

In this work, we collect a field version of an existing lab dataset and quantify the differences between the two. We examine both the raw differences in hold and latency times and also the impact of these differences on an evaluation of a keystroke-dynamics anomaly detector. Along the way, we lay out the methodology by which researchers can compare field and lab versions of a dataset.

The rest of this thesis proceeds as follows. Part I contains a description of the system that was constructed to collect the field dataset and describes our data collection methodology. Part II contains the analysis of the meta-data surrounding the two datasets. Part III contains the analysis of the raw typing data. Finally, in Part IV we quantify the differences between anomaly-detector evaluations conducted on field and lab data.

Chapter 2

Related Work

We are unaware of previous work that compares data sets gathered under field vs. lab conditions. Therefore, this chapter will introduce keystroke dynamics in general, and speak to the matter of why the difference between field and lab data might be an issue.

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. Much recent work has 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 among 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 among 21 subjects using a Support Vector Machine. It is noteworthy that when comparing keystroke algorithms on a fair basis (e.g., using the same data set), uncomplicated techniques turn out to work better than the more complex ones; for example, Killourhy and Maxion [2009b] showed that a classifier based on simple 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, each of whom composed messages of roughly 65 characters, while Samura and Nishimura [2009] obtained 100% accuracy in discriminating among 52 subjects who composed Japanese free text. A recent survey of the field was conducted by Teh et al. [2013].

Many keystroke studies collect data in rather unprincipled ways. There are few controls on timing, venue, subject population, sampling, number of subjects, stimulus materials, confounds, power, etc. These factors can influence final results, so it is essential that benchmark data sets accommodate them. Many of these same factors can be influenced to greater or lesser extents by whether or not data are collected in lab or field conditions, since lab studies are much more likely to be controlled. Hence, a natural conclusion might be that lab data are superior in quality to field data, but this has never been demonstrated. This thesis addresses that issue.

Part I

Experiment and System

Evaluating differences between lab and field data requires the collection of a field version of an existing lab data set. In this part of the thesis, we will describe the data collection methodology used in our study, and the system implemented to collect a field data set. Chapter 3 gives a brief description of the data collection procedure used in our lab, and explains how this procedure was modified to collect a field data set. The resulting modifications required the construction of a new data collection system. In Chapter 4, we present the design of this system and its operational details.

Chapter 3

Data Collection and Experiment Setup

We now turn to the objectives of the current study, a discussion of the lab dataset already collected, and a list of the adjustments necessary to gather a field version of that dataset.

3.1 Objectives

We will accomplish the following objectives:

- Gather a field version of an existing lab dataset.
- Examine what differences there are between the existing lab dataset and the field dataset.
- Explore the impact of these differences on detectors

A new system will be built to facilitate the collection of the field dataset.

3.2 Previous lab experiment

In previous lab study, we recruited 51 subjects to participate in an experiment spanning across 8 consecutive weekdays; each day involved a typing session consisting of 50 repetitions of the password ".tie5Roanl". Our primary approach to recruit subjects was by word-of-mouth. Before the experiment started, each subject would give consent on paper form and fill out a demographic form consisting of questions about age, gender and factors that might affect typing on paper. A research assistant was involved to distribute these forms, collect subject responses, provide instructions and answer any questions subjects might have. All equipment was prepared in the lab including a desktop with an external keyboard to collect keystroke data. We developed MTP (see section 3.3) to record keystrokes. After each typing session, the research assistant would ask questions on factors that might affect the subject's typing on that day.

3.3 Data collection software – Metri-Text Prompter (MTP)

We developed MTP to record keystroke data. Figure 3.1 shows the interface. The phrase to be typed is displayed in the window. The subject will type the phrase in the text box followed by the Enter key. The subject will repeat this process until 50 repetitions of the phrase have been completed. Whenever the subject makes a typing mistake, the text box will be greyed out and the program will ask the subject to type that repetition again. Once the subject finishes all 50 repetitions, MTP will generate a data file containing the keystroke data.

3.4 Need for a new system

Our current research goal requires us to collect a field version of the existing lab dataset. One way to achieve that goal would be to take our lab machine to a public place and ask people to type. However, since our experiment requires multiple sessions to be finished

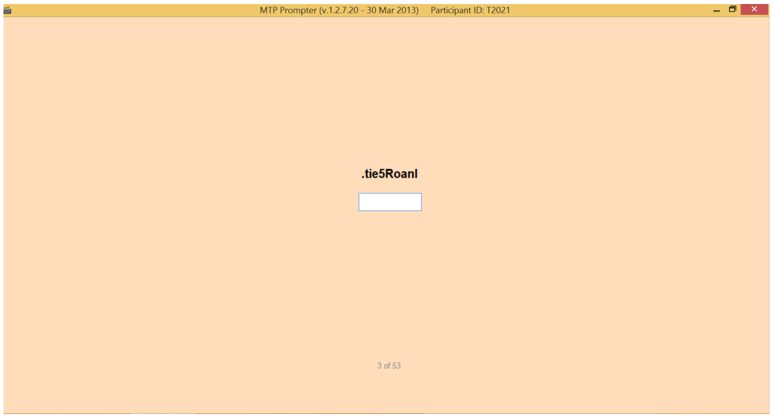

MTP Prompter (v.1.2.7.20 - 30 Mar 2013)   Participant ID: T2021
.tie5Roanl
3 of 53

Figure 3.1: MTP key logger interface.

across multiple days, it would be unrealistic to ask subjects to come to the same place over consecutive days to complete our experiment. Hence, we opted to collect keystroke data over the Internet.

However, MTP is not equipped with a network functionality that can send the typing data to a server. Furthermore, we did not have a server-side program to receive this data. Therefore, we need to design and implement a system to collect keystroke data remotely. We call this new system remote MTP. It is vital that we keep the entire data collection procedure as close as the one used in lab. In total, we want remote MTP to have the following capacities:

1. **Enrollment.** Subjects should be able to give consent electronically and install remote MTP.
2. **Collect demographic information.** The system must be able to administer the same demographic survey and transmit this information back to us.
3. **Collect typing data.** The system must be able to collect typing data and transmit this data back to us.

4. **Post-typing questionnaire.** The system must be able to administer the same post-typing questionnaire after each typing session and send the data back.
5. **Data verification.** The system must be able to detect, report and fix any integrity problems related to typing data.
6. **Maintenance and updating.** The system must provide a way for us to maintain and update its components. Once there is a software update, the system must be able to deliver it to all subjects in the study automatically.

Among all the goals we want the system to achieve, we focus on the following:

1. **Data integrity.** Missing or corrupted data will have adverse effects on data analysis. In the lab study, we would check for missing or corrupted data and the completion of the experiment by every subject, all of which we want to be done automatically by the new system.
2. **Ease of use for subjects.** We want to create a nice and easy-to-use interface for subjects.
3. **Ease of maintenance and update.** We want to be able to maintain and update the system relatively easily.

3.5 Field experiment

We will collect a field version of the existing lab dataset. 51 subjects will be recruited to complete an experiment spanning across 10 consecutive days; each day involves a typing session consisting of 50 repetitions of the password “.tie5Roanl”. We recruit subjects over the Internet, thus removing geographical limitations on subject recruitment. Before the experiment starts, instead of giving consent and filling out demographic form on paper, subjects will complete these forms electronically. After each typing session, subjects will fill out an electronic version of the questionnaire used in the lab study. All the data is collected using remote MTP. Although we collect 10 sessions of data, we only use the first eight sessions in our analysis since we only collected eight sessions from each subject in previous lab study.

Chapter 4

System Architecture for Remote MTP

Remote MTP is a system we built to collect a field version of the existing lab data set. It follows a traditional client-server model where the client side collects and transmits typing data. The server side program receives typing data from the client, verifies its integrity, and handles bookkeeping tasks to facilitate our experiment. We will now step through system operations to explain how it achieves the desired capabilities discussed in Chapter 3 and satisfies our goals on data integrity, ease of use for subjects and ease of maintenance and updating.

4.1 System operations

We recruited subjects through word-of-mouth. Once a subject expressed interest in the experiment, we would email them the link to a welcome page of our study and then begin the enrollment.

4.1.1 Enrollment

Figure 4.1 shows the client and server interactions for enrollment.

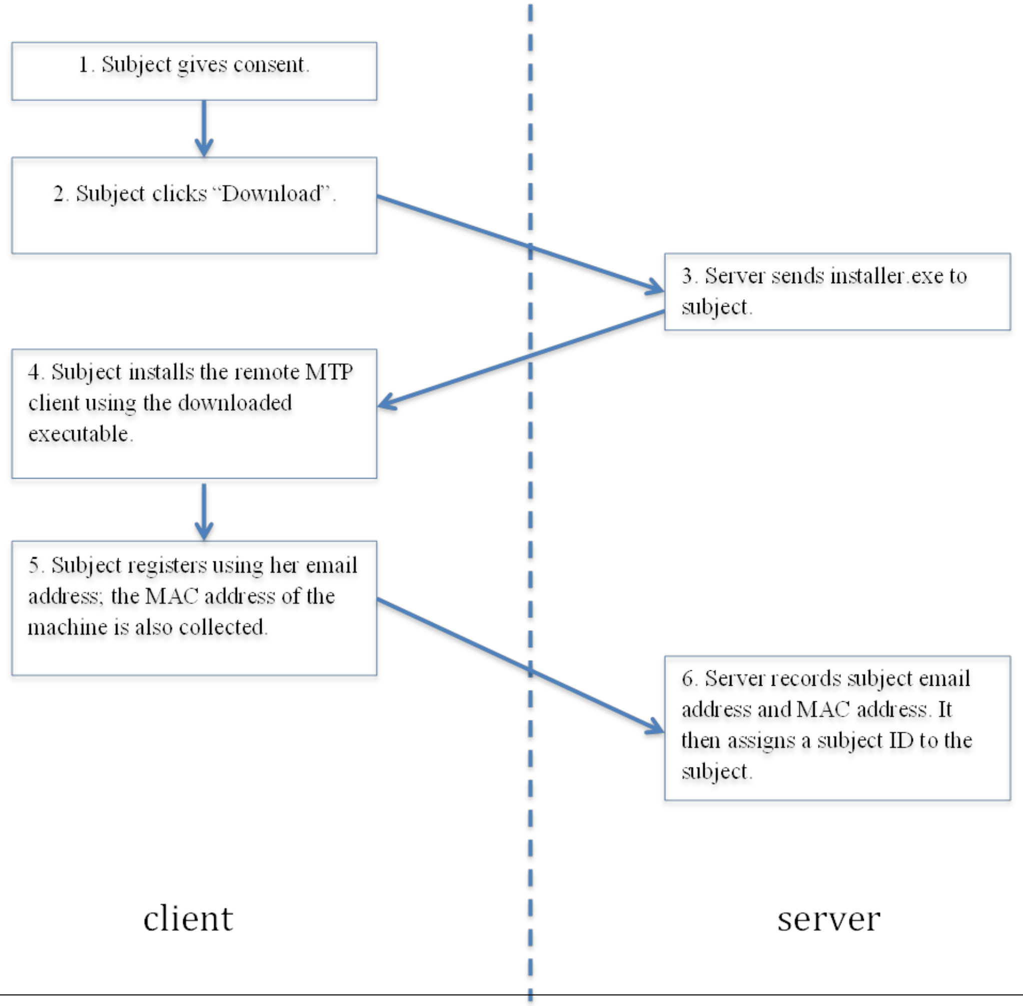
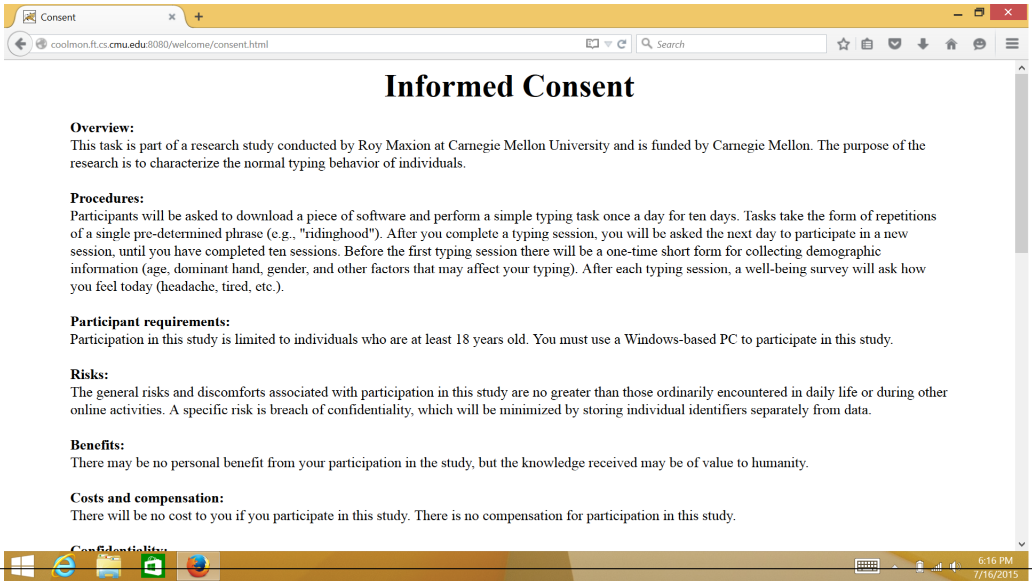

1. Subject gives consent.
2. Subject clicks "Download".
3. Server sends installer.exe to subject.
4. Subject installs the remote MTP client using the downloaded executable.
5. Subject registers using her email address; the MAC address of the machine is also collected.
6. Server records subject email address and MAC address. It then assigns a subject ID to the subject.
client
server

Figure 4.1: Flow chart showing interactions between client and server during enrollment.

Step 1: Subject gives consent.

Once the subject clicks on the link in the e-mail, she will be taken to a welcome page providing a brief overview of what the experiment is and what we expect her to do. The consent form will appear by clicking "Proceed" button, and all listed terms will be agreed when the subject clicks on "Proceed to download" button. Figure 4.2 shows the consent form.



Informed Consent

Overview:
This task is part of a research study conducted by Roy Maxion at Carnegie Mellon University and is funded by Carnegie Mellon. The purpose of the research is to characterize the normal typing behavior of individuals.

Procedures:
Participants will be asked to download a piece of software and perform a simple typing task once a day for ten days. Tasks take the form of repetitions of a single pre-determined phrase (e.g., "ridinghood"). After you complete a typing session, you will be asked the next day to participate in a new session, until you have completed ten sessions. Before the first typing session there will be a one-time short form for collecting demographic information (age, dominant hand, gender, and other factors that may affect your typing). After each typing session, a well-being survey will ask how you feel today (headache, tired, etc.).

Participant requirements:
Participation in this study is limited to individuals who are at least 18 years old. You must use a Windows-based PC to participate in this study.

Risks:
The general risks and discomforts associated with participation in this study are no greater than those ordinarily encountered in daily life or during other online activities. A specific risk is breach of confidentiality, which will be minimized by storing individual identifiers separately from data.

Benefits:
There may be no personal benefit from your participation in the study, but the knowledge received may be of value to humanity.

Costs and compensation:
There will be no cost to you if you participate in this study. There is no compensation for participation in this study.

Confidentiality:

Figure 4.2: Web page containing the consent form.

Step 2: Subject clicks “download”.

The subject will be directed to the download page, providing instructions on how to install the remote MTP client and how to complete the registration. The download starts after the subject clicks on “download” button.

Step 3: Server sends installer executable to subject.

Step 4: Subject installs the remote MTP client using the downloaded executable.

The downloaded executable is a traditional Windows application installer and requires only a few clicks of the “next” button to finish. Instructions are provided on the download page in case subjects are confused. Once installation is finished, the remote MTP client will automatically start.

Step 5: Subject registers using her email address; the MAC address of the machine is also collected.

The remote MTP client will prompt the subject to enter an email address during registration and require the subject to enter her email address twice to confirm correctness since we might contact the subject later in the experiment.

We also collect the MAC address in the background and sent it to server along with the collected email address.

The MAC address is used as an identifier of the subject's keyboard and acts as the safeguard against the use of multiple keyboards in the experiment since changing keyboards or machines in the middle may affect keystroke data in unknown and unpredictable ways.

Step 6: Server records subject email address and MAC address. It then assigns a subject ID to the subject.

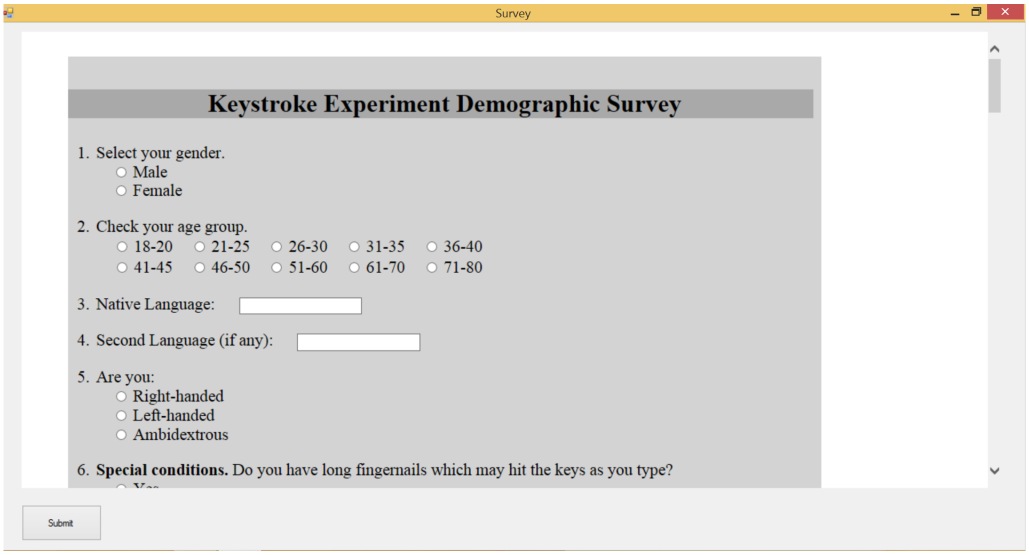
The MTP server will try to search through the database for a duplicate using the received email address. If a duplicate is found, then the subject has already registered and thus the server will return the corresponding subject ID. Otherwise, it will assign the next available subject ID.

Having two or more subject IDs being assigned to the same subject can have potential damage to data integrity. The mapping from email address to subject ID acts as a safeguard against this potential danger.

4.1.2 Collect demographic information

Step 1: Subject fills out demographic form.

The remote MTP client will display an electronic version of the demographic form used in previous lab study and require the subject to fill it out. Once the subject finishes filling out the form, she will click the “Submit” button. The remote MTP client will then send this information to the server. Figure 4.3 shows the demographic form.



Keystroke Experiment Demographic Survey

1. Select your gender.
   ○ Male
   ○ Female

2. Check your age group.
   ○ 18-20 ○ 21-25 ○ 26-30 ○ 31-35 ○ 36-40
   ○ 41-45 ○ 46-50 ○ 51-60 ○ 61-70 ○ 71-80

3. Native Language: ______

4. Second Language (if any): ______

5. Are you:
   ○ Right-handed
   ○ Left-handed
   ○ Ambidextrous

6. **Special conditions.** Do you have long fingernails which may hit the keys as you type?

Figure 4.3: Remote MTP client displaying the demographic form.

Step 2: Server records subject responses.

4.1.3 Collect typing data

Figure 4.4 shows the client and server interactions during collection of typing data.

Step 1: Remote MTP client prompts the subject for a typing sample. Subject clicks “Yes”.

The remote MTP client prompts the subject immediately after enrollment, when the machine reboots or when 20 hours have passed since the last typing session. Figure 4.5 shows the interface of remote MTP client. The subject can start a typing session by clicking the

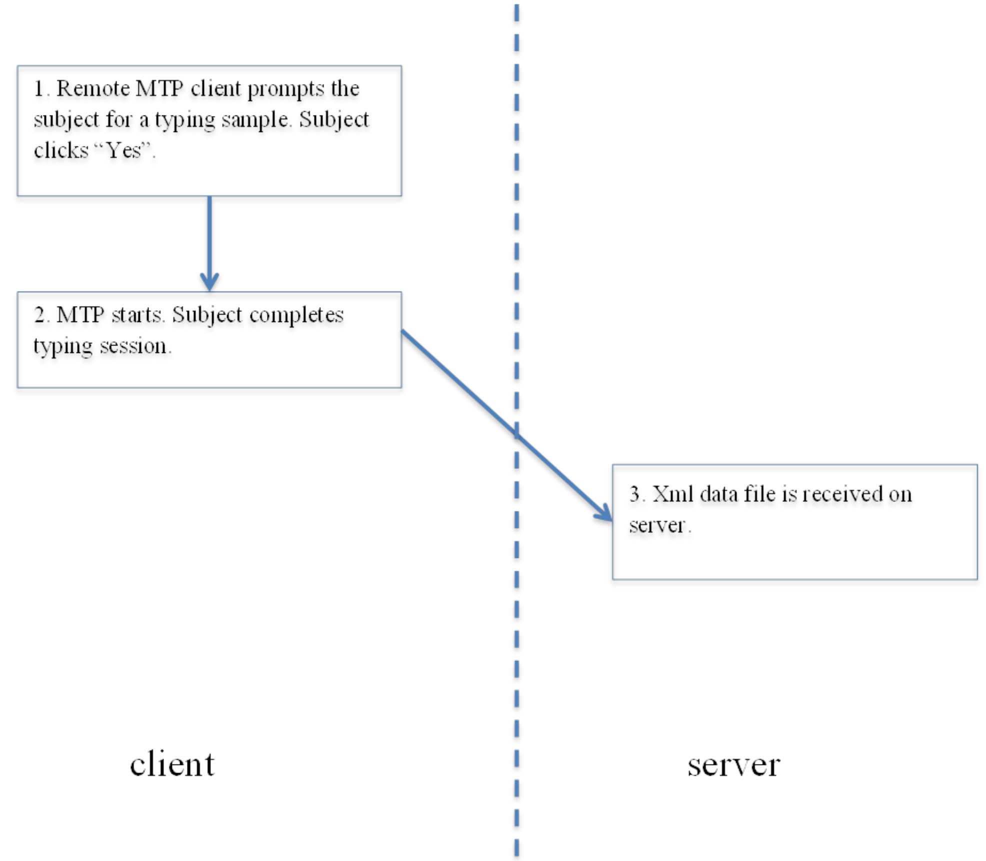
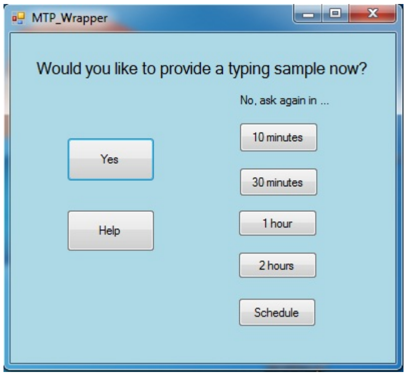

1. Remote MTP client prompts the subject for a typing sample. Subject clicks "Yes".
2. MTP starts. Subject completes typing session.
3. Xml data file is received on server.
client
server

Figure 4.4: Flow chart showing the interactions between client and server during the collection of typing data.

"Yes" button. If the subject is busy at the moment, she can click one of the time buttons to wait for the specified period. For example, if the subject wants to be prompted again in two hours, she can click the "2 hours" button. By default, there will be a waiting period of twenty hours between two typing sessions. However, the subject can set up a custom schedule for the application by clicking the "Schedule" button. The remote MTP client will then prompt the subject at scheduled times.

Step 2: MTP starts. Subject completes typing session.



Would you like to provide a typing sample now?

Figure 4.5: Interface for scheduling and starting a typing session.

MTP appears and the subject will start typing. An XML file containing typing data will be generated once the session is complete and this file is sent to the MTP server. If a subject quits MTP before finishing the typing session, the XML file will be marked as “Incomplete” by prepending the “Incomplete” string to the file name. This incomplete information is required for data verification on the MTP server. In case a connection to server cannot be established, the client will keep trying until the connection is established.

Step 3: XML data file is received by the server.

Once the MTP server receives the typing data file from the remote MTP client, it will place the file in directories corresponding to the subject and experiment pair.

4.1.4 Post-typing questionnaire

Step 1: Subject fills out the post-typing questionnaire.

After the subject finishes typing, the remote MTP client will display an electronic version of the questionnaire used in previous lab study. After the subject finishes filling out the questionnaire, the remote MTP client will send the information back to the server in the background.

Step 2: Questionnaire responses are received on server.

4.1.5 Data verification

Step 1: XML data files received by MTP server are checked for integrity .

The MTP server will assign one of four labels to each XML data file based on the result of an integrity check:

- **Incomplete file**: A data file that contains data for an incomplete typing session. Remote MTP client prepends "incomplete" string to the file name if the session is terminated early as discussed before.

- **Complete file**: A data file that contains data for 50 repetitions of ".tie5Roanl".

- **Recoverable file**: A data file with some missing repetitions due to common artifacts in the XML file which can be easily fixed. For example, one common artifact is the hitting of backspace. Subjects tend to press backspace when they make a typo. The backspace keystroke is then recorded at the beginning of the next repetition and causes problems. The fix is to simply remove the backspace keystroke from the data file.

- **Suspicious file**: Any file that does not belong to the above three categories and thus needs further investigation by researchers. One common cause is the hitting of keys outside of the standard character set. For example, some subjects had euro keys in their keyboards and might accidentally press those keys during typing. The fix is to simply remove these keystrokes from the data file.

Step 2: Server records progress.

The MTP server will update the database according to the label of the received file. Each complete file or recoverable file received counts towards the subject's final progress. Once the subject reaches our expected number of sessions, the MTP server will mark that subject as finished. A suspicious file will only be counted if a researcher can manually fix its problem. Otherwise it will be ignored. Incomplete files are ignored.

4.1.6 Maintenance and updating

From time to time, we wish to deploy bug fixes or need to modify the questionnaire form. Hence, it is important for the client side application to have the ability to download updates. The remote MTP client checks for updates in the following three situations:

1. The remote MTP client starts and the subject still has not completed his set of typing sessions.
2. The subject finishes a typing session.
3. The subject is assigned more typing sessions by the researchers.

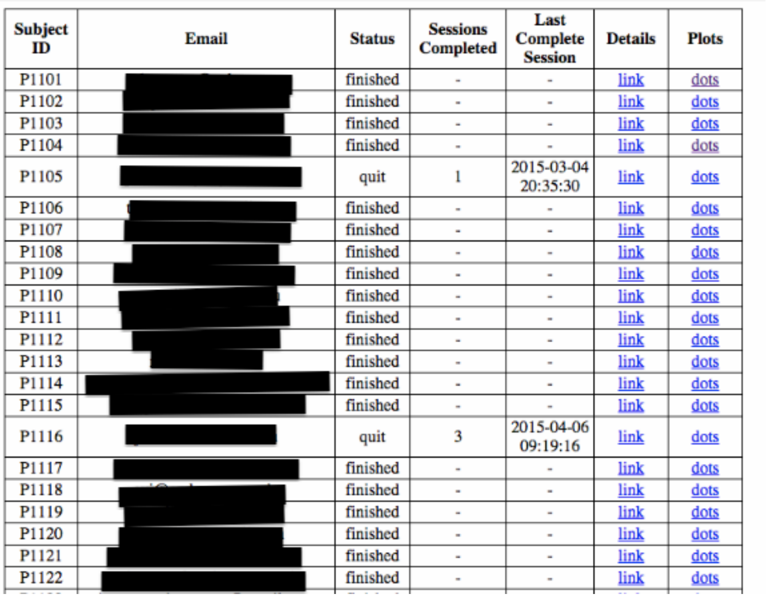
In any of the above three situations, the remote MTP client will check whether it has the latest version. If it does not, it will contact the MTP server for an update package. In case a connection to server cannot be established, the remote MTP client will try one more time and then wait until one of the above three situations occurs.

4.2 Experience Running the System

In general, the system worked well during data collection. However, we encountered a few noteworthy problems.

4.2.1 Status web page

During our study, we found it hard to visualize the keystroke data collected. We thus developed a web page containing information for each participating subject. Figure 4.6 shows the web page. The link under "Plots" point to a page containing images visualizing the collected typing data for each session of that subject.



Subject ID	Email	Status	Sessions Completed	Last Complete Session	Details	Plots
P1101	[redacted]	finished	-	-	link	dots
P1102	[redacted]	finished	-	-	link	dots
P1103	[redacted]	finished	-	-	link	dots
P1104	[redacted]	finished	-	-	link	dots
P1105	[redacted]	quit	1	2015-03-04 20:35:30	link	dots
P1106	[redacted]	finished	-	-	link	dots
P1107	[redacted]	finished	-	-	link	dots
P1108	[redacted]	finished	-	-	link	dots
P1109	[redacted]	finished	-	-	link	dots
P1110	[redacted]	finished	-	-	link	dots
P1111	[redacted]	finished	-	-	link	dots
P1112	[redacted]	finished	-	-	link	dots
P1113	[redacted]	finished	-	-	link	dots
P1114	[redacted]	finished	-	-	link	dots
P1115	[redacted]	finished	-	-	link	dots
P1116	[redacted]	quit	3	2015-04-06 09:19:16	link	dots
P1117	[redacted]	finished	-	-	link	dots
P1118	[redacted]	finished	-	-	link	dots
P1119	[redacted]	finished	-	-	link	dots
P1120	[redacted]	finished	-	-	link	dots
P1121	[redacted]	finished	-	-	link	dots
P1122	[redacted]	finished	-	-	link	dots

Figure 4.6: Figure showing the status web page.

4.2.2 Anti-virus software

One recurring problem is related to anti-virus software. This is due to the nature of MTP program. MTP is a key-logger that only records keystrokes when it is active. However, there are still some anti-virus software prevents MTP from running by terminating MTP whenever it starts. When this occurs, we usually asked the subject to disable the anti-virus software during the 10-day period.

4.2.3 Machine freeze

Some subjects had the experience that their machines froze in the middle of a typing session. When a machine freezes while the subject is typing, the typed characters will not show up until one or two seconds later. The digram-latency time or hold-time for the affected features will become significantly longer than average. The freezing of a machine might be due to the running of other applications on the machine at the same time. For our current study, we treat these frozen sessions as regular sessions.

Part II

MetaData Analysis

Meta-data refers to data that describes the field and lab keystroke data sets that we have collected. Its primary value lies in its explanatory capabilities; differences between the field and lab data sets may be partly or fully explained by differences in meta-data. In this part of the thesis, we search for such differences, focusing on demographic information.

Chapter 5

Demographics Analysis

5.1 Research question

Are there distributional differences between lab and field subjects in terms of gender, age, handedness and type of keyboard used?

5.2 Motivation

Keystroke typing rhythms can be heavily influenced by demographic factors. We have chosen to focus on four factors that are likely to have a large impact: gender, age, handedness, and the type of keyboard used.

For example, men tend to have shorter digram-latency time since men generally have bigger hands compared to women and thus can cover wider area of a keyboard. Young people are more familiar with computer technology and thus are more likely to be faster typists. If subjects of one data set are mainly composed of young people, then we would expect the average total typing time of that data set to be shorter. Keys typed by the dominant hand may have shorter digram-latency time. Left-handed subjects will have a different set of keys typed by the dominant hand as compared to right-handed subjects, resulting in a different combination of keys with shorter digram-latency time. Some keys might be farther apart in one type of keyboard than another, causing an increase in digram-latency time.

5.3 Method

Information on gender, age, handedness and the type of keyboards was collected using online questionnaires at the beginning of the experiment for field subjects. Lab subjects filled out the same form on paper before they started the experiment. Subjects were asked to state their sex (male & female), their age group (10 choices ranging from 18 to 80+), handedness (right-handed, left-handed and ambidextrous), and the type of keyboard they were currently using (choices are shown in Table 5.1).

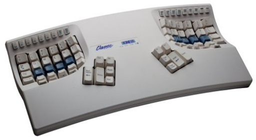
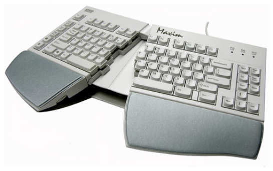
standard	laptop	natural
		
kinesis	kinesis maxim	mac external

Table 5.1: Types of keyboards used in the field study

We analyzed the demographic data in terms of these four attributes. For each attribute, we obtained subject counts for each sub-category and compared them between the lab and field data sets.

5.4 Results

Gender. Table 5.2 shows the gender summary for lab and field subjects. Lab and field subjects are quite similar in terms of gender ratio. The male-female ratio is about 3:2 for both data sets.

Gender	Count (Field)	Count (Lab)
Female	20	21
Male	31	30

Table 5.2: Gender summary for lab and field data.

Age. Table 5.3 shows the subject counts of each age group for lab and field subjects. The lab subjects are more evenly distributed among age groups. Over half of the field subjects (28 out of 51 subjects) came from the 21-25 age group. However, only 10 out of 51 lab subjects came from the same age group. Both groups did not have subjects from the 71-80+ age group.

Age Group	Count (Field)	Count (Lab)
18 - 20	1	5
21 - 25	28	10
26 - 30	2	9
31 - 35	1	3
36 - 40	2	4
41 - 45	3	3
46 - 50	3	7
51 - 60	6	8
61 - 70	5	2
71 - 80+	0	0

Table 5.3: Counts of subjects in each age group for lab and field data.

Handedness. Table 5.4 shows the handedness summary for lab and field subjects. Lab and field data sets have roughly the same number of subjects in each category. The ratio between right-handedness and left-handedness is about 8:1 for both data sets.

Handedness	Count (Field)	Count (Lab)
Right-handed	45	44
Left-handed	6	7

Table 5.4: Handedness summary for lab and field data.

Types of keyboard used. Table 5.5 shows the subject counts for the different keyboard types used in the field study. We noticed that field subjects using standard keyboards and laptop keyboards were the majority, with each keyboard type being used by 23 subjects. There were no subjects using either kinesis keyboard or kinesis maxim keyboards. Three subjects used a natural keyboard. Two subjects used a mac external keyboard. All lab subjects used the same standard keyboard.

Keyboard Type	Count (Field)	Count (Lab)
standard	23	51
laptop	23	0
natural	3	0
kinesis	0	0
kinesis maxim	0	0
mac external	2	0

Table 5.5: Counts of subjects using each kind of keyboard for lab and field data.

5.5 Discussion

We noticed that the field data set did not reflect the general populace as well as the lab data set. More than half of our field subjects were between 21 and 25 years of age, a much higher proportion than in the general population. The field data set consisted of mainly college students because we started asking our friends when recruiting subjects and many of them were still in college. When the recruitment was carried out later by word of mouth, it was mainly among the student population.

The young-people-dominated population in field study may explain some of the differences in keystroke data between lab and field data sets. For example, if the field subjects can type faster than the lab subjects on average, then one possible explanation is that field data set consists of mainly young people who are better typists.

Moreover, both data sets do not reflect the true gender ratio. According to the 2010 census in the United States, the male-female ratio is roughly 1:1 (49.2:50.8) [Cen, 2011]. Comparing with the gender ratio from lab and field data sets, we had more male subjects in both data sets.

5.6 Summary

In this section, we examined differences between lab and field subjects on gender, age, handedness and types of keyboard used. The findings can be summarized as follows:

- Lab and field subjects are similar in terms of gender ratio. Both have a male-female ratio of 3:2.
- Field subjects are more concentrated around the 21-25 age group while lab subjects are more evenly distributed among all age groups.
- Lab and field subjects are similar in terms of handedness. Both have a ratio of 8:1 between right-handedness and left-handedness.
- Most field subjects used standard and laptop keyboards. Lab subjects all used the standard keyboard in our lab.

Part III

Typing Data Analysis

In this part of the thesis, we examine the differences between the typing in the lab and field datasets. As we analyze these differences, we take into consideration that subjects' typing will change as they gain practice at typing the password. In Chapter 6, we start our analysis by examining differences in central tendency and spread in the two datasets for both unpracticed and practiced typing data. Chapter 7 provides a more in-depth examination of the differences in distribution, with a focus on verifying common distributional assumptions used in keystroke dynamics detectors. Chapters 8 and 9 contain an exploration of the differences in rate of practice, in terms of the total typing time and individual keystroke features, respectively. This exploration sheds some light on how data collection and evaluation procedures could be made more efficient. Finally, we explore data variability in Chapter 10, due to the significant influence of variability on detector performance.

Chapter 6

Characterization by Descriptive Statistics

6.1 Research Question

How are lab and field data different in terms of their characterizations via the basic descriptive statistics offered by a simple 5-number summary – min, median, max, and quartiles?

6.2 Motivation

We begin our analysis of lab and field data by offering a basic statistical description of the data in terms of simple central tendency and spread. These provide a general sense of the fundamental differences between our two data sets such that patterns might emerge to provide guidance on how to further characterize and distinguish these data sets. This will provide context for the rest of our investigations. Both data sets, lab and field, comprise eight sessions of 50 typing repetitions each. We examine here only sessions 1 and 8, because session 1 is very unpracticed, and session 8 is very well practiced, and we expect to find different characterizations with respect to low and high practice.

6.3 Method

We began by generating the five-number summaries for Session 1 (unpracticed) and Session 8 (practiced) total typing time for both lab and field datasets. The total typing time is the total amount of time to finish typing one password. We then generated box-plots for each feature from Session 1 and Session 8 for both data sets. In addition, we generated bar charts to compare median values of individual features between lab and field data set for Session 1 and Session 8. Finally, we computed the inter-quartile ranges for individual features by computing the difference between the first and third quartiles and generated bar charts. Five-number summaries of individual features (min, first quartile, median, third quartile, max) for lab and field data can be found in Appendix A.

6.4 Results

Lab vs. field. Table 6.1 shows the five-number summaries for Sessions 1 and 8, total typing time, computed from lab and field datasets. In Session 1 (unpracticed) the values for min, 1st quartile and median have large differences between lab and field data, – 32%, 12% and 9.1% respectively. On the other hand, also for Session 1, the values for the 3rd quartile and the max show little difference between lab and field – 3.3% and 1.5% respectively.

Regarding Session 8 (highly practiced), only the min and max values show large differences between lab and field data – 19.5% and 126.2% respectively. The values for the 1st quartile, median and 3rd quartile values show little change between lab and field data.

Session 1 vs. Session 8. As we move from Session 1 (unpracticed) to Session 8 (highly practiced) we see that the lab and field data start to look more and more similar. For example, the lab/field difference in the 1st quartile of Session 1 was 12%, and that same difference in Session 8 was 5.3%, for a reduction of more than half as practice accumulates. There are similar reductions in the values of the median and 3rd quartile.

Session 1

Venue	Min	1st quartile	Median	3rd quartile	Max
Lab	1.295	2.271	2.801	3.610	36.000
Field	0.880	1.997	2.544	3.490	35.450

Session 8

Venue	Min	1st quartile	Median	3rd quartile	Max
Lab	1.083	1.692	2.077	2.635	9.720
Field	0.872	1.603	1.937	2.592	21.980

Table 6.1: Five-number summaries, total typing time in seconds, lab vs. field data, for Sessions 1 (unpracticed) and 8 (highly practiced). As described in the text, there are large differences between lab and field data. However, these differences diminish as we move from Session 1 (unpracticed) to Session 8 (highly practiced).

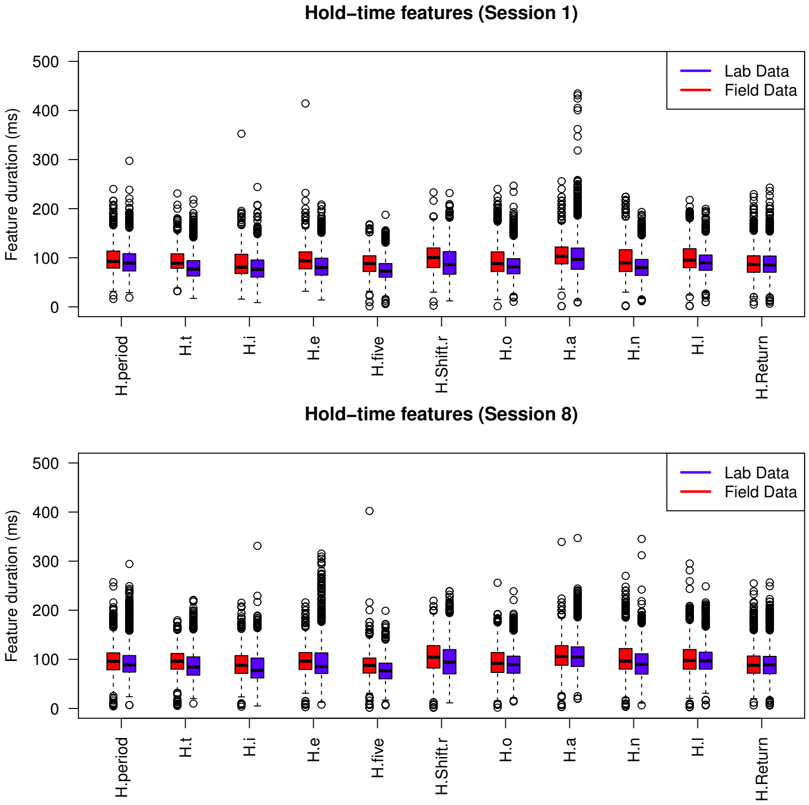


Figure 6.1: **Box plots for hold-time features.** Features from field dataset are pictured in red while those from lab dataset are pictured in blue. The three lines in each box from bottom to top denotes: 1st quartile, median and 3rd quartile. The dots denote outliers that are outside 1.5 inter-quartile range. Note that the percentages of outliers in Session 1 hold-time feature data for lab and field are 3.3% and 3.6%. The percentages of outliers in Session 8 hold-time feature data for lab and field are 3.2% and 4.0%.

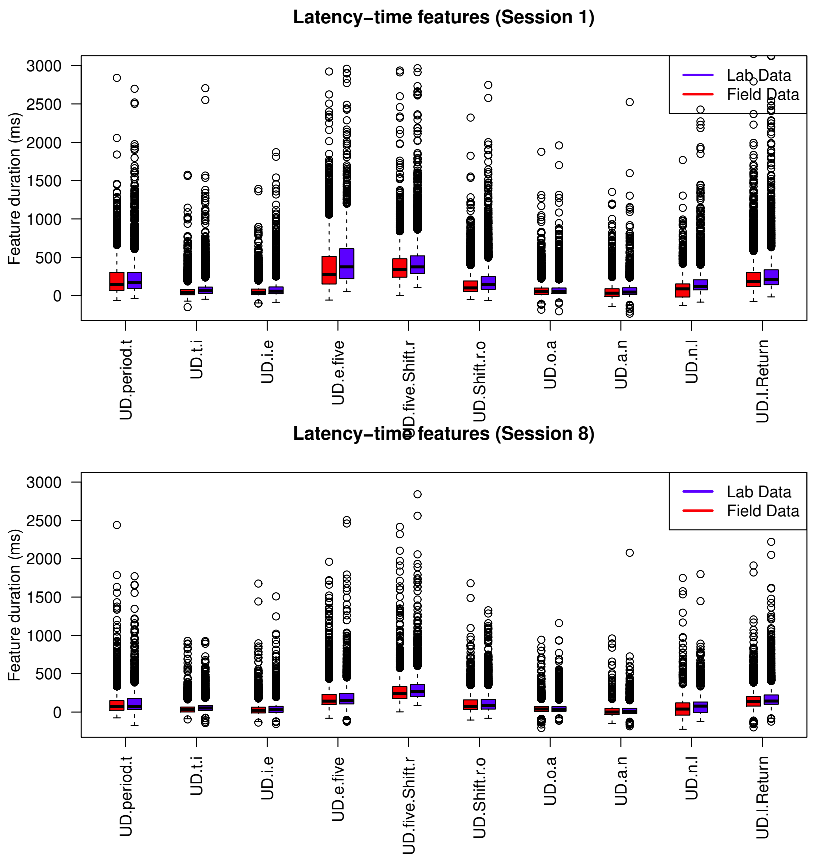


Figure 6.2: **Box plots for latency-time features.** Features from field dataset are pictured in red while those from lab dataset are pictured in blue. The three lines in each box from bottom to top denotes: 1st quartile, median and 3rd quartile. The dots denote outliers that are outside 1.5 inter-quartile range. Note that the percentage of outliers in Session 1 latency-time feature data for lab and field are 8.1% and 7.8%. The percentage of outliers in Session 8 latency-time feature data for lab and field are 7.0% and 6.6%.

Figures 6.1 and 6.2 show box plots for hold-time and latency-time features from Session 1 and Session 8 for both lab (blue) and field (red) data.

Each box represents three aspects of the data. The bottom of the box denotes the 1st quartile; the top of the box denotes the 3rd quartile; and the horizontal line inside the box denotes the median (2nd quartile). The interquartile range (IQR), also called the midspread or middle fifty, is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles. The vertical dashed lines are sometimes called whiskers; their ends represent the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile. The open circles above and below the whiskers denote outliers that are outside the 1.5 inter-quartile range.

In general, the box plots show that feature data from the field is similar to that from the lab. Despite differences in spread and center shown in the plots, they generally follow a similar trend. Specifically, features with greater spread in lab data also have greater spread in field data, and vice versa.

The boxplots in both Figures 6.1 and 6.2 show a fair number of outliers, but in fact the outliers are of less concern than the boxplots make them appear. For the hold-time features shown in Figure 6.1, the proportion of outliers in the data is quite small – 3.3% and 3.6%, respectively, for lab and field data in the unpracticed condition (Session 1). Similarly, as shown in Figure 6.2, the proportion of outliers is small – 8.1% lab and 7.8% field for latency times in practiced data (Session 8). It is natural that latency times show higher outlier rates than hold times do, because hold times are less under conscious psychomotor control.

The boxplots show a similarity between the lab and field data on a scale of only 100-500 milliseconds. However, many features have shorter durations, and it can be difficult to observe differences in values of medians and inter-quartile ranges in boxplots. Bar charts, on the other hand, can provide us a higher resolution and thus show these differences more clearly. These bar charts are presented below.

Bar charts - for a finer look at the data. The bar charts in this section provide a finer-grained examination of the data. Here we compare the medians for the different data sets, as well as the inter-quartile ranges.

Figure 6.3 compares lab and field median values for hold-time features in Sessions 1 (low practice) and 8 (high practice). In Session 1, the biggest difference in median values between lab and field data is 15.4 milliseconds for H.five. The smallest difference is 1.1 milliseconds for H.Return.

In Session 8, the biggest difference in median values is 11.8 milliseconds for H.t. The smallest difference is 0.75 milliseconds for H.l. Note that the median values for hold-time features from both sessions of field data are mostly greater than those in lab data – which is to say that in the figure the red bars are generally taller than the blue bars. One exception is H.Return in Session 8, but the difference is only 1 millisecond ($\sim 1\%$ of the median).

Figure 6.4 shows similar comparisons, but for latency-time features. In Session 1, the largest difference in median values is 98.9 milliseconds for UD.e.five; this is not unexpected, because the transition from the e-key to the 5-key is awkward, and typically takes a long time. The smallest difference is 5.7 milliseconds for UD.o.a; again, this is not surprising, because the o-key is typically struck with a finger on the right hand, and the a-key is typically struck with a finger on the left hand – when opposing hands strike consecutive keys, it can be very fast. In Session 8 (high practice), the biggest difference in median values is 38.55 milliseconds for UD.n.l. The smallest difference is 2.8 milliseconds for UD.o.a. Again, these make sense, because of the fingerings that are typically very fast. Note that the median values for latency features from both sessions of field data are generally smaller than those of lab data. One exception is UD.o.a from Session 8, showing field data has a greater median than lab data.

Figure 6.5 shows compares the inter-quartile range in lab and field data for Sessions 1 and 8, hold-time features. In Session 1, 7 out of 11 hold-time features have a greater inter-quartile range in the field data. In Session 8, 4 out of 11 hold-time features have a greater inter-quartile range in the field data. As subjects become more familiar with the password, hold-time feature data in field becomes less variable.

Figure 6.6 shows the inter-quartile range comparison between lab and field data for Sessions 1 and 8 latency-time features. In Session 1, 5 out of 10 latency-time features have a greater inter-quartile range in the field data. In Session 8, 4 out of 10 latency-time features have a greater inter-quartile range in the field data. Note that field and lab latency-time data show similar variability in both sessions.

6.5 Discussion

This section has provided simple statistical characterizations of the data, intended more for benchmarking against other data sets than for eliciting patterns. Some of the observations have reasonably obvious explanations, but not all. It is surprising that hold-time features in field data have greater median values than in lab data. This may be due to subjects having more practice on their own (field) keyboards, as it is known that higher practice levels lead to longer hold times. Moreover, it is interesting to observe that latency-time feature data in the lab has greater median values than in field data; this may be due to subjects being less familiar with (less practiced on) the keyboard used in the lab.

6.6 Summary

In this chapter, we address the question: how are lab and field data different in terms of central tendency and spread? The findings can be summarized as follows:

- Lab and field data become more similar from Session 1 to Session 8.
- Field data has higher median values for most hold-time features. Lab data has higher median values for most latency-time features.
- Field data has 7 out of 11 hold-time features with greater inter-quartile range in Session 1, but the number diminishes (4 out of 11) in Session 8. Field data has 5 out of 10 latency-time features with greater inter-quartile range in Session 1, but that number also diminishes to 4 in Session 8.
- Variability in the data, either lab or field, as shown by inter-quartile range, shows no clear pattern. There is a mixture of features with greater or lesser variability between lab and field data, and between unpracticed and highly practiced data.

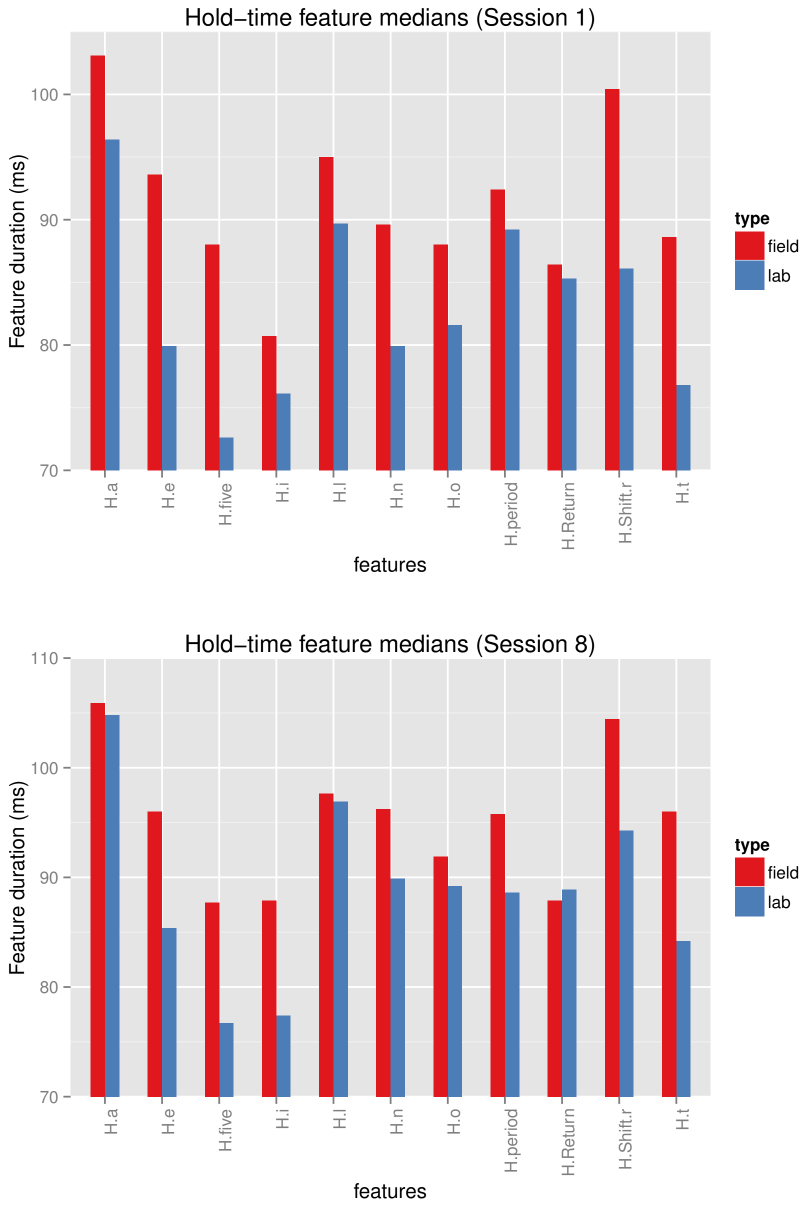


Figure 6.3: Median hold-time features, lab (blue) vs. field (red) data for Session 1 (top) and Session 8 (bottom). In Session 1, the largest difference is 15.4 milliseconds for H.five while the smallest difference is 1.1 milliseconds for H.Return. In Session 8, the largest difference is 11.8 milliseconds for H.t, while the smallest difference is 0.75 milliseconds for H.l. Field data generally have higher medians.

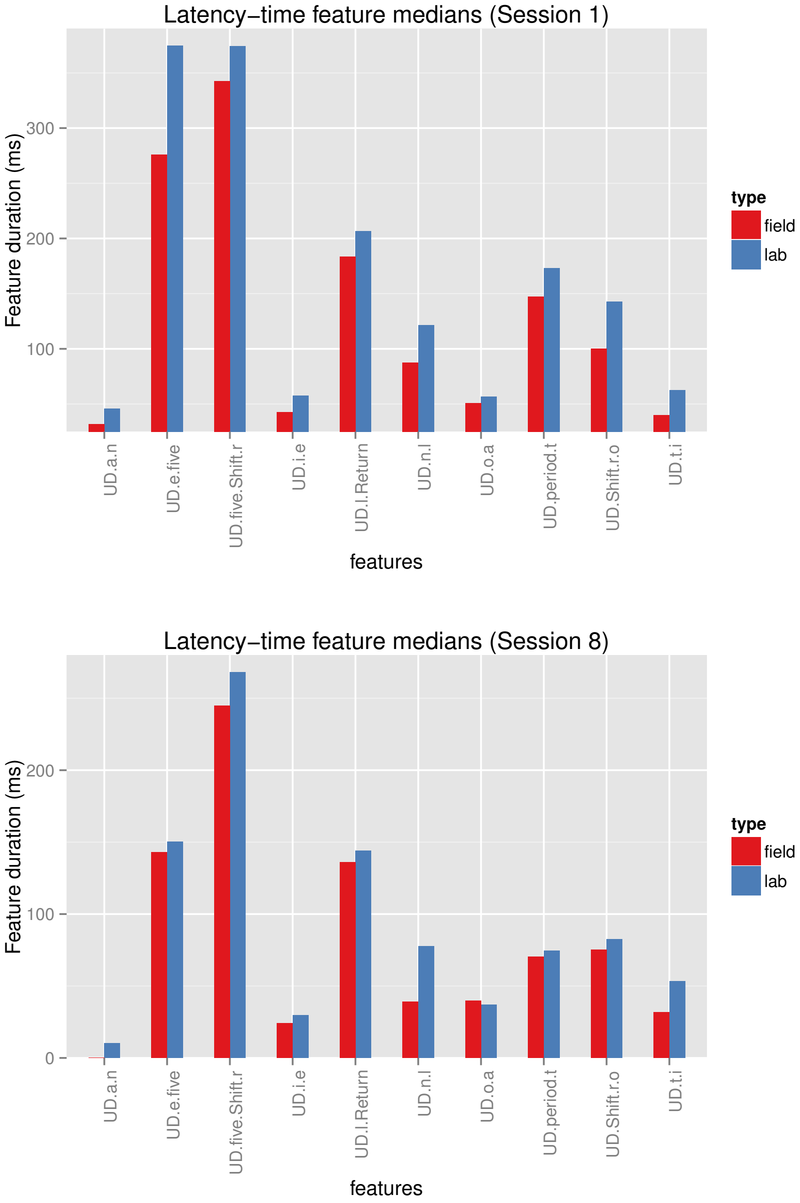


Figure 6.4: Median latency features, lab (blue) vs. field (red) data for Session 1 (top) and Session 8 (bottom). Note that in Session 1, the largest difference is 98.9 ms for UD.e.five; smallest is 5.7 ms for UD.o.a. In Session 8, the largest difference is 38.55 ms for UD.n.l, while the smallest is 2.8 ms for UD.o.a. Latencies are longer in lab data, perhaps due to subjects being less used to the keyboard.

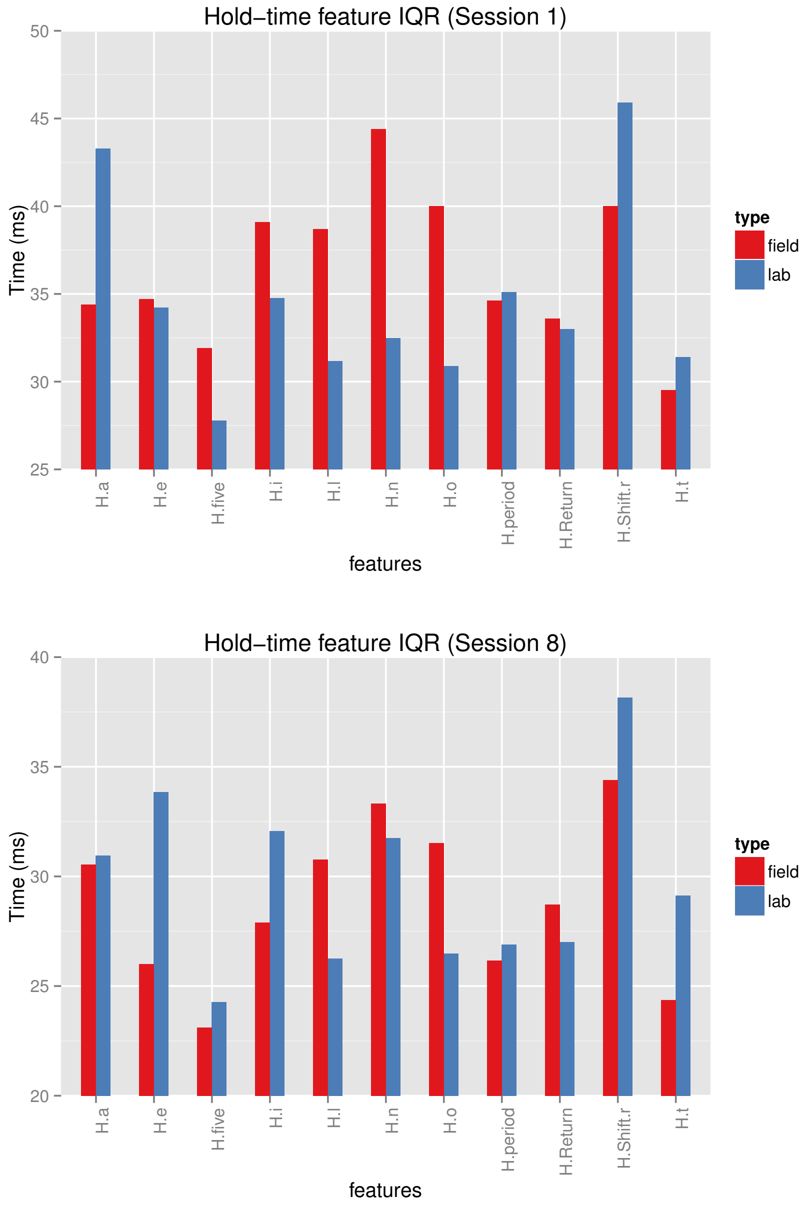


Figure 6.5: Inter-quartile range comparison between lab (blue) and field (red)data for Session 1 (top) and Session 8 (bottom) hold-time features. Note that in Session 1, 7 out of 11 hold-time features have a greater inter-quartile range in the field data. In Session 8, this drops to 4 out of 11. No pattern is apparent.

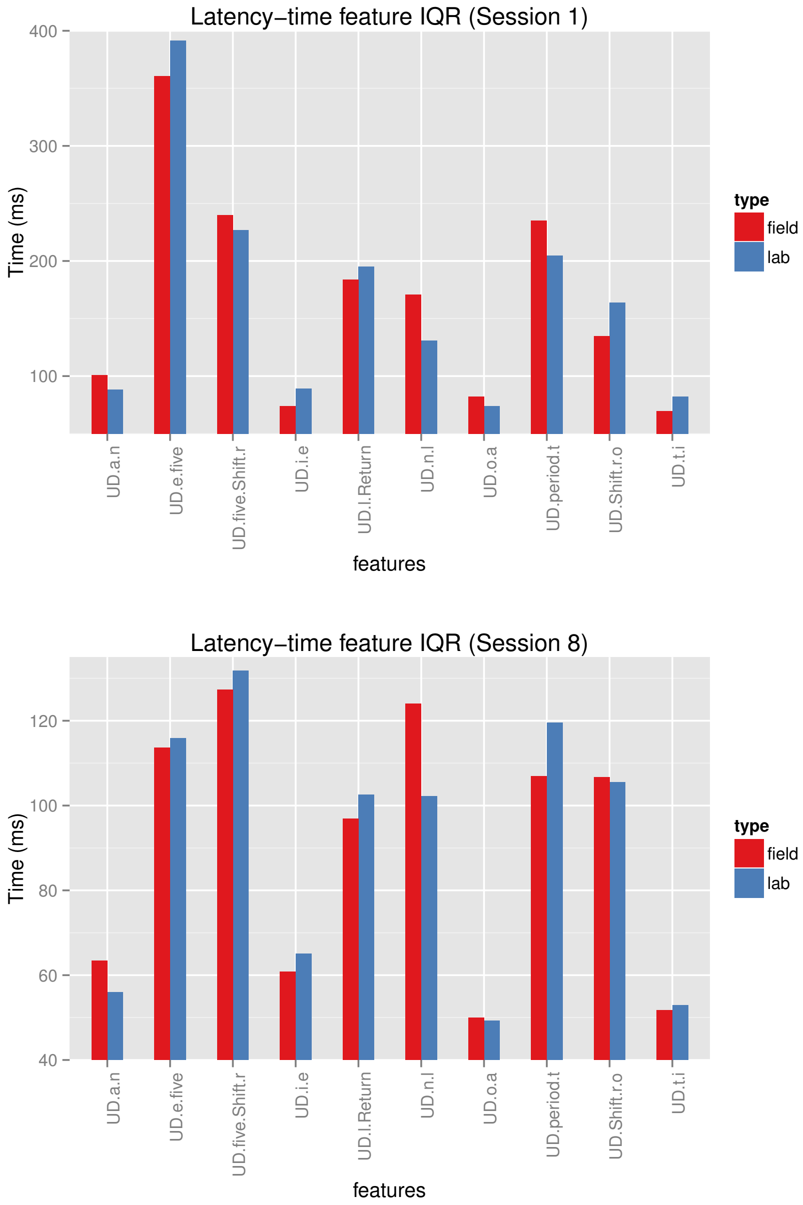


Figure 6.6: Inter-quartile range comparison between lab (blue) and field (red)data for Session 1 (top) and Session 8 (bottom) latency-time features. Note that in Session 1, 5 out of 10 latency-time features have a greater inter-quartile range in the field data. In Session 8, this drops to 4 out of 10 with no overall pattern being apparent.

Chapter 7

Distributions

7.1 Research Question

Are the total typing times of field and lab data similarly distributed? Specifically, are they both normally distributed? Also, how different are their distributions?

7.2 Motivation

In keystroke dynamics, assumptions are often made about the underlying distribution of keystroke data. For instance, detectors may be designed to work on lab data before being transitioned to the field. If field data comes from a different distribution than lab data, then these detectors may not work properly. A particularly common assumption is that the data are normally distributed. If this assumption is violated, keystroke detectors may work significantly worse in the field than in the lab.

7.3 Method

We now turn to the tools used in our analysis.

Quantile-quantile plot (QQ plot). A quantile-quantile plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other [Wilk and Gnanadesikan, 1968]. For a point (x, y) on the plot, x and y correspond to the same quantile of both distributions. If the two distributions are similar, then the points in the plot will fall approximately on the line y = x.

Skewness and kurtosis. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point [NIST, 2012]. It is defined by the following formula

$$\gamma_1 = \mu_3/\mu_2^{3/2} \tag{7.1}$$

where μ_2 and μ_3 are the second and third central moments. The k^{th} central moment of a data population is defined as

$$\mu_k = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^k \tag{7.2}$$

where x_i is the i^{th} data point and N is the number of data points in the population.

Before we dive deep into skewness, we will introduce the concept of a tail. Tail refers to the part of a distribution that is far away from the mean. Left tail refers to the part of a distribution that is far below the mean while right tail refers to the part that is far greater than the mean.

Positive skewness indicates that data are skewed right, meaning the right tail is long relative to the left tail. Negative skewness indicates that data are skewed left, meaning the left tail is long relative to the right tail.

Kurtosis describes the shape of a probability distribution. It is defined as

$$\gamma_2 = \mu_4/\mu_2^2 - 3 \tag{7.3}$$

where μ_2 and μ_4 are the second and fourth central moments.

For a symmetric unimodal distribution, positive kurtosis indicates heavy tails and peakedness relative to the normal distribution, whereas negative kurtosis indicates light tails and flatness [DeCarlo, 1997]. For a normal distribution, the skewness is zero and the kurtosis is three.

Effect size. In statistics, an effect size is a quantitative measure of the strength of a phenomenon [Kelley and Preacher, 2012]. In our study, we will use Cohen's d effect size to measure the magnitude of the difference between total typing times of lab and field data sets. Cohen's d effect size has three categories: small (0.0 $\leq$ d $\leq$ 0.2), medium (0.2 $<$ d $<$ 0.8), and large (d $\geq$ 0.8) [Cohen, 1992].

We now turn to the data analysis. In this study, we focus on the total typing time. Since total typing time is the sum of individual features and thus a summary statistic, differences in the distribution of the total typing time would imply differences in the distributions of individual features. Moreover, differences in total typing time are generally results of changes in a combination of features and thus can provide us a context for the micro variations in typing data.

To calculate the total typing time of one repetition, we add up all hold times and keyup-keydown times in that repetition. For both lab and field data sets, we computed the total typing time for every repetition in each session for all subjects. In the end, we obtained two groups of total typing time data, one for lab and one for field, each containing 51 subjects $\times$ 8 sessions per subject $\times$ 50 repetitions per session = 20400 total typing times.

To begin with, we checked the normality of the total typing time data. We first plotted histograms for the two groups of total typing time data and compared them with estimated normal curves. After that, we generated normal QQ plots for both groups. We also computed the skewness and kurtosis for both groups.

We then started to see whether the total typing time data for lab and field come from the same distribution. We first generated quantile-quantile plot for the two groups of data to provide visual evidence. We then conducted Mann-Whitney U test with the null hypothesis being that two samples come from the same population. Finally, we computed Cohen's d effect size to quantify the difference of total typing time data between lab and field data sets.

7.4 Results

7.4.1 Normality of total typing time data

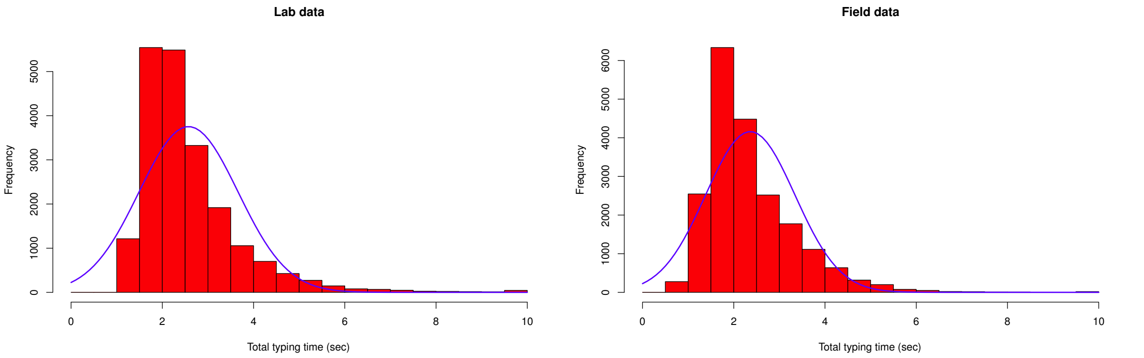


Figure 7.1: Histograms for total typing time with overlaid normal distribution curves estimated using sample mean and sample standard deviation.

Figure 7.1 shows the histograms of total typing time for both lab and field data. We also extracted the sample mean and the sample standard deviation of the total typing time for both groups and plotted the estimated normal distribution curves (colored in blue) using them. Although both histograms have a nice bell shape and resemble the estimated normal distribution curves, both histograms show a high peak near the mean, showing a deviation from normal distribution.

Since these histograms do not give us enough evidence to verify normality of the total typing time data, we generated quantile-quantile plots. For data coming from a normal distribution, the points plotted in the QQ plot should fall approximately on the line y = x. Figure 7.2 shows QQ plots for total typing time computed from lab and field data sets. Data points in both plots do not fall on the line y = x, indicating a deviation from normality.

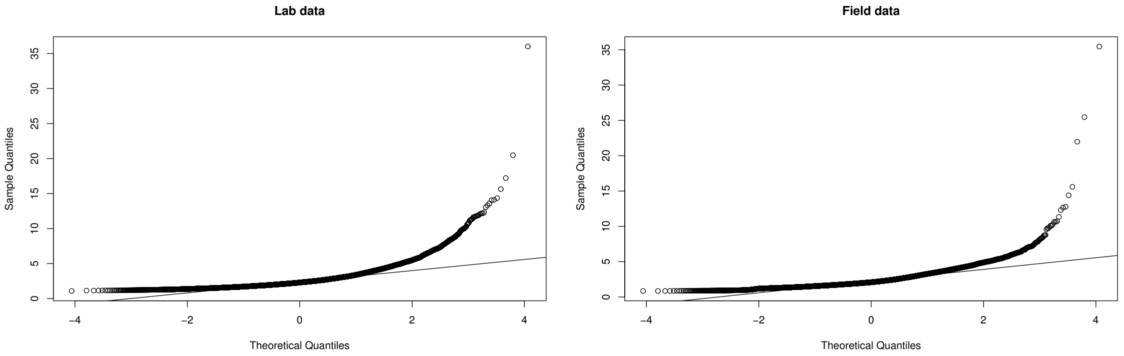


Figure 7.2: QQ plots for total typing time computed from lab and field data sets respectively.

To further confirm the evidence from QQ plot, we calculated the skewness and kurtosis for the total typing time data from both lab and field data sets:

- Skewness for lab data: 3.92
- Kurtosis for lab data: 54.6
- Skewness for field data: 4.19
- Kurtosis for field data: 80.13

Skewness values for both lab and field data show that the total typing time data are skewed right, meaning that the right tail is long relative to the left tail. Kurtosis values for both lab and field data show that the total typing time data has heavier tails and is more peaked relative to the normal distribution. We can see them reflected in Figure 7.1. The skewness and kurtosis values are way larger than those for a normal distribution, showing a deviation from normality.

7.4.2 Statistical Difference

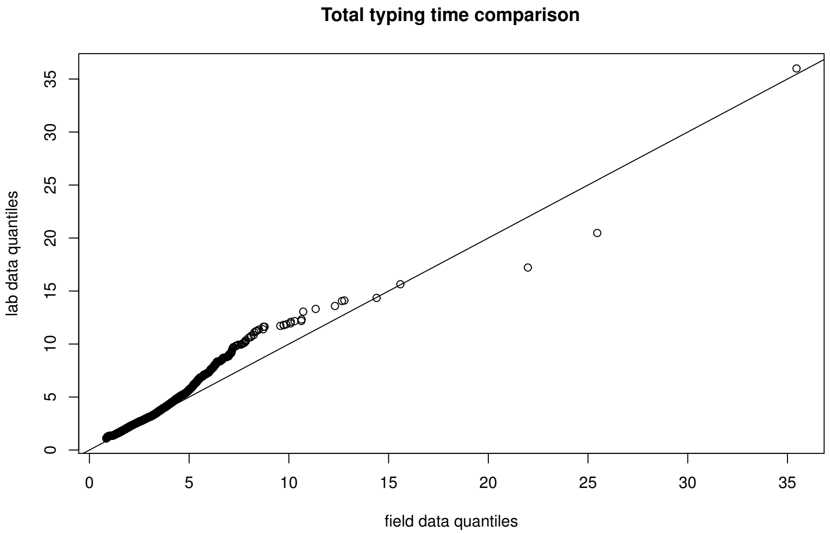


Figure 7.3: Quantile-quantile plot for total typing time computed from lab and field data sets.

Figure 7.3 shows the quantile-quantile plot of total typing time data computed from lab and field data sets. The straight line from the origin only goes through the first half of the quantile-quantile plot. It provides evidence that the total typing time data computed from the lab data set does not come from the same distribution as that computed from the field data set.

Since the total typing time data cannot be modeled using normal distribution, we chose to use Mann-Whitney U test which has a greater efficiency than the unpaired t-test on non-normal distributions. The null hypothesis is that two samples come from the same population. The alternative hypothesis is that two samples come from distinct populations. We used a significance level of α = 0.05. The p-value obtained is less than 2.2e-16 (W = 178564436). Since the p-value is below 0.05, we reject the null hypothesis. Therefore, the total typing time computed from the lab and field data set come from distinct distributions. In addition, we computed the Cohen's d effect size to quantify the difference of total typing time between lab and field data sets. The effect size is 0.197, indicating small difference between the means of the two groups.

7.5 Discussion

Although the non-parametric t-test shows lab data and field data are statistically different in terms of total typing time, Cohen's d effect size shows that the magnitude of difference between the two groups is small.

In our study, the sample is quite large, consisting of 20400 data points. A significance test depends on both sample size and effect size. With a sufficiently large sample, even the smallest difference between two data sets can result in significant results. Hence, it is possible that the statistically significant result we obtained is due to the large sample size. However, effect size is independent of sample size. Therefore, we think that the Cohen's d effect size is more informative than the result from the significance test. In this case, it is better to say that while there are differences in the total typing time between lab and field data but these differences are relatively small as measured by Cohen's d effect size.

7.6 Summary

In this section, we answer the question: are the total typing times of field and lab data similarly distributed? Specifically, are they both normally distributed? Also, how different are their distributions? The findings can be summarized as follows:

- Lab and field data are not normally distributed in terms of total typing time.
- There are significant differences between lab and field data in terms of total typing time. However, the differences are small as indicated by a Cohen's d effect size of 0.197.

Chapter 8

Practice Curves

8.1 Research Question

How differently does practice influence total typing time between lab and field datasets? Specifically, does the total typing time become stable at different points? Also, is the total typing time roughly the same after practice for both datasets?

8.2 Motivation

Practice can have a large impact on keystroke dynamics data. Differences in the practice effect between lab and field data could lead researchers to change their data collection and data evaluation procedures.

For example, if the average lab subject becomes practiced around repetition 300, and we require 100 repetitions of practiced typing data to train and test the detector, then we need 400 repetitions from a lab subject (300 to attain high practice, and an additional 100 for training and testing). On the other hand, if the average field subject becomes practiced around repetition 200, then we only need 300 repetitions from a field subject (200 to attain high practice, and an additional 100 for training and testing). The total number of repetitions diminishes by 100 in the field study. Evaluations of detector performance will be more meaningful when subjects from different datasets have the same level of practice.

8.3 Method

We now turn to the tools used in our analysis. We used the concept of practice curves and stabilization points from Lau and Maxion [2015].

Practice curve. In our study, a subject will type a password 50 times per day for 8 days. During the 8-day period, the subject will become more and more practiced at typing the password. This will be reflected in the reduced total typing time from day 1 to day 8. We plotted the total typing time for all repetitions of the subject as data points. A practice curve is a curve fitted to all the data points. We fitted a full power-law learning curve to the data. The power-law learning curve resembles an exponential curve. As the number of repetitions increases, the rate of change becomes smaller and smaller until the curve becomes relatively flat. The formula for the practice is

$$y = M + B \cdot (x + E)^{-\beta} \tag{8.1}$$

where y is the predicted amount of time (in seconds) it will take to type the x^{th} repetition of the password. Here we give a description of what the four parameters are in the formula:

M: the predicted minimum amount of time (on average) required for the subject to type the password, even with an infinite number of practice repetitions.

B: the "range of learning" which predicts the total difference (in seconds) in the total typing time, between the fully unpracticed state and the fully (infinitely) practiced state.

E: the prior experience of a subject before s/he began our typing task, cast in terms of the number of repetitions s/he has already typed.

β: the learning parameter, which governs how quickly a subject learns.

The parameters are chosen to minimize the mean absolute deviation of the residuals.

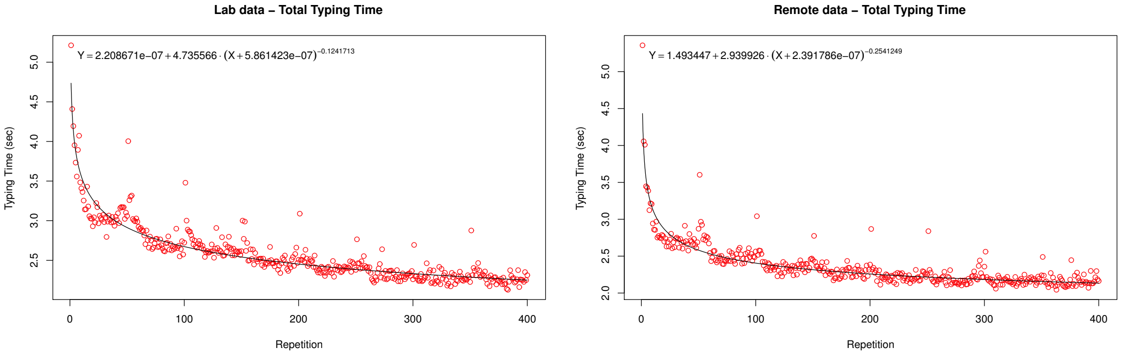
Stabilization point. In our study, we used stabilization points to quantify the level of practice. The 1% and 0.1 % stabilization points are the points where the average time to type the next repetition of the password decreases by 1% and 0.1%. The 1 ms stabilization point is the point where the average time to type the next repetition of the password decreases by one millisecond. A subject who has reached the 1 ms stabilization point is considered to be fully practiced.

We now turn to the data analysis. In this study, we focus on the practice effect on total typing time. For each dataset, we computed the average total typing time for each repetition over all subjects in the dataset. Hence, for a dataset, we have 400 data points (50 repetitions × 8 sessions). These data points reflect the total typing time change of an average subject. We then fitted a full power-law learning curve to the 400 data points.

After we obtained the learning curves for both lab and field data, we used them to make predictions on the total typing time beyond the existing 400 repetitions. We used these predictions to show what the total typing time would be after subjects become fully practiced. We compared these predictions between lab and field subjects. We also computed the 1%, 0.1% and 1 ms stabilization points. We compared these stabilization points to see how quickly subjects from lab and field datasets became practiced.

8.4 Results

Figure 8.1 shows the practice curves fitted to lab and field data respectively.



(a) Lab practice curve

(b) Field practice curve

Figure 8.1: Lab vs. Field. Full power-law learning curve fitted to the total typing time of 400 repetitions of the password .tie5Roanl. The figure on the left side shows the practice curve for the average lab subject. The figure on the right side shows the practice curve for the average field subject. Red circles mark single instances of the typed password.

The power-law equation for the practice curve fitted to lab data is

$$Y = 2.208671\text{E-}07 + 4.735566 \cdot (X + 5.861423\text{E-}07)^{-0.1241713} \tag{8.2}$$

The equation shows that the average lab subject is able to type the password in 2.209E-07 seconds (M) after an infinite amount of practice. The predicted improvement is 4.74 seconds (B) over the fully unpracticed state. The learning rate is 0.12 (β). The prior-experience value (E) is 5.86E-07, showing no experience in typing the password.

The power-law equation for the practice curve fitted to field data is:

$$Y = 1.493447 + 2.939926 \cdot (X + 2.391786\text{E-}07)^{-0.2541249} \tag{8.3}$$

The equation shows that the average field subject is able to type the password in 1.49 seconds (M) after an infinite amount of practice. The predicted improvement is 2.94 seconds (B) over the fully unpracticed state. The learning rate is 0.25 (β). The prior-experience value (E) is 2.39E-07, showing no experience in typing the password.

Findings: The average field subject became practiced more quickly. Based on the two practice curve equations, we can see that the average field subject has a larger learning parameter indicating that the average field subject gets practiced more quickly than the average lab subject does.

Notice that the average lab subject can finish typing the password in almost 0 (2.208671E-07) second after infinite amount of practice. We felt this was unrealistic and therefore used the two power-law equations to predict total typing time at various points up to repetition 1000. Table 8.1 shows the results.

Repetition	Prediction for Lab	Prediction for Field	Difference
500	2.19	2.10	-0.09
600	2.14	2.07	-0.07
700	2.10	2.05	-0.05
800	2.06	2.03	-0.03
900	2.04	2.02	-0.02
1000	2.01	2.00	-0.01

Table 8.1: Total typing time (sec.) prediction for the average subject from lab and field.

Findings: The total typing time of practiced lab and field subjects differs by at most 4% between repetitions 500 - 1000. All predicted total typing time of the average field subject is less than those of the average lab subject. As we increased the repetition number, the differences became smaller and smaller. Compared to the total typing time, the difference is around 4% or less. Hence, the total typing time after practice for the average lab subject and the average field subject is about the same.

Using the two power-law equations, we computed the stabilization points for the average subjects. Table 8.2 shows three stabilization points we computed for both lab and field datasets. For example, for the average lab subject, at repetition 13, the average time to type the password changes by only 1% or 34.4 ms per repetition. This means that the next repetition will take 34.4 milliseconds less.

Repetition	Absolute change in ms (lab)	Repetition	Absolute change in ms (field)
13	34.4 (1%)	14	28.6 (1%)
125	2.6 (0.1 %)	98	2.4 (0.1 %)
292	1 (0.04 %)	197	1 (0.04 %)

Table 8.2: Stabilization points indicating the improvement of the average subject's typing for lab data (the left side) and field data (the right side) due to practice.

Findings: The typing data for the average field subject stabilized more quickly. We can see from Table 8.2 that the typing of the average field subject became stable more quickly than the average lab subjects did. The average field subject becomes fully practiced at repetition 197 which is about 100 repetitions less than the average lab subject.

8.5 Discussion

We found that the average field subject became practiced more quickly than the average lab subject. We wonder whether the familiarity of keyboards might be the cause. Since the field subjects used their own keyboards, they were more familiar with the keyboards and were likely to take less time to get practiced. On the other hand, for the lab subjects, they had to first get acquainted to the keyboard used in our lab and might take longer for their typing data to become stable.

8.6 Summary

This section answers the question: how differently does practice influence total typing time between lab and field datasets? Specifically, does the total typing time become stable at different points? Also, is the total typing time roughly the same after practice for both datasets? The findings can be summarized as:

- The average field subject got practiced faster than the average lab subject.

- The total typing time for the average field subject after practice is within 4% of the total typing time for the average lab subject.

We hypothesize that field subjects became practiced much faster than lab subjects due to their familiarity with the keyboard used.

Chapter 9

Practice Effect on Features

9.1 Research Question

How differently does practice influence individual features for lab and field data sets?

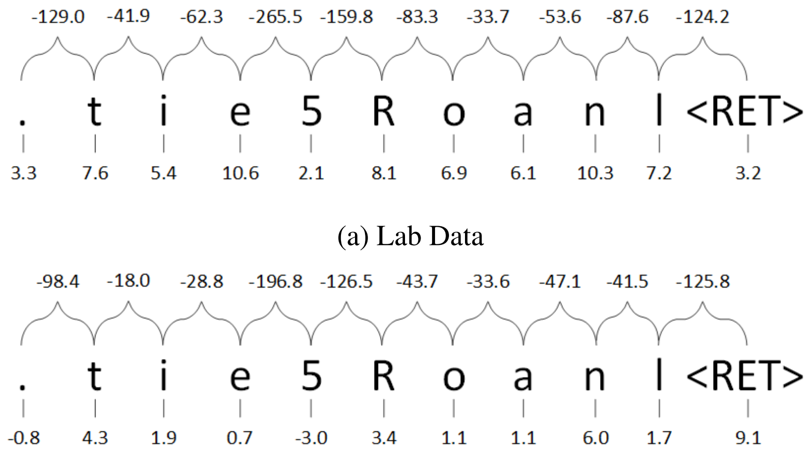
9.2 Motivation

Changes in individual features (e.g., hold and latency times) due to practice are important characteristics of typing data. Models used by anomaly detectors capture these changes and utilize them in classifications. Similar trends in features are often assumed in field data. However, if this assumption is violated, anomaly detectors may have worse performance in the field than in the lab.

9.3 Method

We first computed the average, absolute changes in hold and latency times in milliseconds. For each subject in the dataset, we computed the mean for Session 1 and Session 8 data corresponding to each feature over all repetitions of a session. After computing the difference between the two means for each feature, we summed these differences and averaged over all subjects to obtain the average, absolute change for each feature. Similarly, we computed the percentage change for each feature.

9.4 Results



(a) Lab Data

(b) Field Data

Figure 9.1: Average, absolute changes in hold and latency times (milliseconds) between session 1 and session 8. **Upper rank:** latency times. **Lower rank:** hold times. Example: in (a), the time between i and e decreased by 62.3 ms from Session 1 to Session 8; similarly, the hold time for period increased by 3.3 ms.

Figure 9.1 shows how the typing of an average subject from lab and field respectively changes from Session 1 (repetitions 1 - 50) to Session 8 (repetitions 351 - 400). Figure 9.1a (from [Lau and Maxion, 2015]). The numbers in the upper row indicate the average, absolute changes of latency times (keyup-keydown). The numbers in the lower row indicate the average, absolute changes of hold times. A positive number indicates that the feature took more time as the subject became more practiced; a negative number indicates that the feature took less time. For example, in Figure 9.1a, the average e-5 keyup-keydown time decreased by 265.5 ms from Session 1 to Session 8.

We also computed the percentage change of each individual feature for lab and field data sets. Table 9.1 and Table 9.2 show the percentage change in hold-time and latency-time features from both lab and field data. The lab percentage change tables are based on [Lau and Maxion, 2015].

As shown in Table 9.1, the maximum percentage change in lab hold-time features is 12.5% for hold-times on key e and n, while the minimum percentage change is 2.8% for the hold-time on key 5. For the field data, we observe the maximum percentage change in field hold-time features is 10.3% for the hold-time on key Return while the minimum percentage is 0.7% for the hold-time on key e. Additionally, as shown in Table 9.2, the maximum percentage change in lab latency-time features is -67.4% for a-n while the minimum percentage change is -33.4% for 5-R. For the field data, the maximum percentage change in latency-time features is -81.7% for a-n and the minimum percentage change is -26.3%.

We made the following observations:

- With respect to the average, absolute change between sessions 1 and 8, all latency-times had negative changes, meaning that these latencies were shorter in session 8 than in session 1. Similar observations are made in field data. However, the magnitude of changes is smaller in field data. In lab data, the biggest absolute decreases in latency-times were for the following four digrams e-5, .-t, 5-R, and l-Return, which were all greater than 100 milliseconds [Lau and Maxion, 2015]. Similarly, in field data, the biggest absolute decreases were seen in the aforementioned four digrams. These digrams seem to be hard to type for both lab and field subjects.

- With respect to percentage change of latency-time features, both datasets have shown that a-n has the biggest percentage change (shown in Table 9.2).

- With respect to the average, absolute change between sessions 1 and 8, all hold times had positive changes, meaning that hold times were longer in session 8 than in session 1. However, some changes in hold times computed from field data were negative indicating that these hold times were shorter. Specifically, hold times for

Lab Hold-times	
Feature	% change
period	3.6
t	9.4
i	6.8
e	12.5
5	2.8
R	8.9
o	8.2
a	5.9
n	12.5
l	7.9
Return	3.6

Field Hold-times	
Feature	% change
period	-0.8
t	4.6
i	2.1
e	0.7
5	-3.3
R	3.3
o	1.2
a	1.0
n	6.2
l	1.7
Return	10.3

Table 9.1: Percentage change in hold-time features, between Session 1 and Session 8, lab vs. field data, averaged over 51 subjects. The maximum percentage change in lab hold-time features is 12.5% for the hold-times on key e and n. The minimum percentage change in lab hold-time features is 2.8% for the hold-time on key 5. The maximum percentage change in field hold-time features is 10.3% for the hold-time on key Return. The minimum percentage change in filed hold-time features is 0.7% for the hold-time on key e.

period, and 5 were shorter. Most changes in hold times had a smaller magnitude in field data. One exception is the Return key. In lab data, the biggest absolute change was 10.6 milliseconds on e. In field data, the biggest absolute change was 9.1 milliseconds on Return.

Lab latency times	
Feature	% change
period-t	-49.9
t-i	-37.4
i-e	-54.0
e-five	-56.8
5-R	-33.4
R-o	-38.9
o-a	-37.7
a-n	-67.4
n-l	-50.2
l-Return	-39.4

Field latency times	
Feature	% change
period-t	-44.2
t-i	-27.5
i-e	-41.3
e-five	-50.3
5-R	-30.8
R-o	-26.3
o-a	-41.4
a-n	-81.7
n-l	-42.5
l-Return	-44.4

Table 9.2: Percentage change in latency-time features between Session 1 and Session 8, lab vs. field data, averaged over 51 subjects. The maximum percentage change in lab latency times is -67.4% for a-n. The minimum percentage change in lab latency times is -33.4% for 5-R. The maximum percentage change in field latency times is -81.7% for a-n. The minimum percentage change in field latency times is -26.3% for R-o.

9.5 Discussion

We observed similar trends in both hold-time feature changes and latency-time features due to practice. Latency changes were all negative for both lab and field data sets. Changes for most hold-time features were positive for both data sets. However, the exceptions were hold times for period and 5. We also noticed that the changes in field data had smaller magnitudes. We think the reason behind these differences is that field subjects are more familiar with their typing environments, specifically their keyboards.

9.6 Summary

This section answers the question: how differently does practice influence individual features for lab and field data sets? The findings can be summarized as follows:

- Hold-time features for both lab and field data sets showed an increase in time.
- Latency-time features for both lab and field data sets showed a decrease in time.

With respect to practice effects on individual features, lab data is similar to field data.

Chapter 10

Feature Variability

10.1 Research Question

How does the variability of individual features differ between lab and field data once subjects are practiced?

10.2 Motivation

In keystroke dynamics, data variability has a large impact on detector performance. Knowing the differences in data variability between lab and field data sets allows us to predict whether detectors designed and tested on lab data are likely to work in the field. For example, if we know that features in field data have greater variability, then the detector performance may be worse when running on field data.

10.3 Method

For each feature, we computed the standard deviation for Session 8 data corresponding to that feature across all repetitions of Session 8 and all subjects. We then ranked the features in increasing order, based on standard deviation.

10.4 Results

Feature Name	Standard Deviation (Lab Data)	Feature Name	Standard Deviation (Field Data)
H.5	0.0222	H.5	0.0259
H.l	0.0284	H.e	0.0305
H.o	0.0285	H.l	0.0330
H.i	0.0298	H.i	0.0341
H.t	0.0299	H.R	0.0351
H.Return	0.0300	H.n	0.0390
H.n	0.0309	H.o	0.0428
H.period	0.0332	H.t	0.0486
H.a	0.0346	H.period	0.0487
H.R	0.0362	H.a	0.0786
H.e	0.0376	H.Return	0.2108

Table 10.1: Lab vs. Field Data. Standard deviation of each hold-time feature from Session 8 for lab (left) and field (right) data. The features are ranked by standard deviation; the least-variable features are at the top and the most-variable features are at the bottom.

Table 10.1 shows the standard deviation of each hold-time feature from Session 8 for lab and field data. We can see that H.5 is the least-variable feature in both data sets. The lab and field data have three out of the four least-variable features in common: H.5, H.l, and H.i. For the other features, the rankings are quite different. After comparing each individual hold-time feature, we can see that most features have greater standard deviation in field data. The exceptions are hold-times on key e, R and Return.

Table 10.2 shows the standard deviation of each digram-latency-time feature from Session 8. The features are ranked by increasing standard deviation. For the lab data, the least-variable digram-latency-time feature is UD.o.a. For the field data, the least-variable digram-latency-time feature is UD.a.n. The lab and field data have all five least-variable

Feature Name	Standard Deviation	Feature Name	Standard Deviation
	(Lab Data)		(Field Data)
UD.o.a	0.089	UD.a.n	0.076
UD.t.i	0.090	UD.o.a	0.080
UD.a.n	0.090	UD.t.i	0.087
UD.i.e	0.110	UD.i.e	0.102
UD.n.l	0.128	UD.n.l	0.139
UD.Shift.r.o	0.156	UD.l.Return	0.146
UD.period.t	0.167	UD.period.t	0.184
UD.l.Return	0.177	UD.five.Shift.r	0.199
UD.e.five	0.198	UD.e.five	0.201
UD.five.Shift.r	0.235	UD.Shift.r.o	0.373

Table 10.2: Lab vs. Field. Standard deviation of each keyup-keydown digram feature from Session 8. The features are ranked by standard deviation; the least-variable features are at the top and the most-variable features are at the bottom.

digram-latency-time features in common. For the other features, the rankings are quite different. After comparing each individual digram-latency-time feature, we can see that four features showed greater standard deviation in field data, hence greater overall variation.

10.5 Discussion

Although the rankings of features based on variability for lab and field data sets appear to be quite different, the set of least-variable features for lab data does not differ that much from that for field data. Knowing the set of least-variable features suggests that detectors operating on those features may be the easiest to transfer from lab to field without significant performance difference.

Lab data is less variable in terms of hold-time features. On the other hand, field data has less variability in terms of digram-latency-time features. Detectors operating on hold-time features are likely to have worse performance on field data due to the increase in variation, as compared to lab data. On the other hand, detectors operating on digram-latency-time features may have better performance on field data for similar reasons.

10.6 Summary

In this section, we answer the question: how does the variability of individual features differ between lab and field data once the subjects are practiced? The findings can be summarized as follows:

- The rankings for hold-time features based on variability for lab and field data are quite different. However, if we look at the top four least-variable features for both lab and field data, we see that they share three common features: H.5, H.l and H.i. Lab feature data has less variability in terms of hold-time features.
- The rankings for digram-latency-time features based on variability for lab and field data are quite different. However, if we look at the top five least-variable features for both lab and field data, they are identical for both data sets. Field feature data is less variable in terms of digram-latency-time features.

Part IV

Effects on detector performance

In Part III of this thesis, we demonstrated that there are significant differences between lab and field data. Here in Part IV, we investigate the effects of these differences on keystroke anomaly-detector performance. Chapter 11 presents performance results from fourteen different detectors operating on lab and field data. Chapter 12 examines how the differences in practice effect in lab and field data affect detector performance. Together, these findings shed light on how detectors designed with lab data will fare on field data.

Chapter 11

Differences in Detector Performance

11.1 Research Question

What is the difference in anomaly detector performance when operating on lab data versus field data?

11.2 Motivation

A prominent application of keystroke dynamics is the analysis of typing rhythms for discriminating among users. When designing a new anomaly detector for this application, researchers typically collect data in a controlled lab environment, develop an anomaly detector that performs well on the collected lab data, and then deploy this detector in the field. An underlying assumption here is that there will be no performance change when the detector runs on field data. As we have pointed out in previous chapters, field data is different from lab data in various ways. We would like to find out whether these differences have any impact on detector performance. If there is not much difference in detector performance when operating on field data, then the detector can be safely deployed in the field. Otherwise, conducting field studies before deploying a detector is crucial.

11.3 Method

We now consider the detectors and tools used in our analysis which are borrowed from [Killourhy and Maxion, 2009a]. We briefly summarize the key aspects here, and direct the reader to the original paper for the full details.

Detectors. In our study, we focus on one class of techniques called anomaly detection. The typing samples of a single, genuine user are used to build a model of the user's typing behavior (training). When a new typing sample is presented, the detector tests the sample's similarity to the model, and outputs an anomaly score. When the anomaly score is above a certain threshold, we label the sample as anomalous. In this study, we compare 14 anomaly detectors running on lab and field data. Some examples are the Manhattan detector using the Manhattan distance measure, the neural-network detector, and the SVM detector. These detectors were chosen to cover several common detector types. We direct the reader to [Killourhy and Maxion, 2009a] for the full detector descriptions.

Detector Performance Measure. In keystroke dynamics, there are two common measures: miss rate and false-alarm rate. Miss rate is the percentage of impostor password attempts that are not detected. False-alarm rate is the percentage of genuine user password attempts that are mistakenly detected as impostors. In this study, we used two performance measures: equal-error rate and zero-miss false-alarm rate. Equal-error rate is the value where miss rate and false-alarm rate are equal. Zero-miss false-alarm rate is the false-alarm rate when miss rate is zero. For both measures, the smaller the value, the better the performance of the detector.

Training and Testing. Briefly, for each detector, we started by selecting a genuine user, and using the rest of the 51 subjects as impostors. During the training phase, we used the first 200 repetitions typed by the genuine user to train the detector. During the test phase, we ran the detector on the remaining 200 repetitions to generate user scores. After that, we ran the detector on the first five repetitions from each impostor to generate impostor scores. We then repeated this process by assigning each subject as the genuine user. An equal-error rate and a zero-miss false-alarm rate were computed for each subject, which we averaged to form a system equal-error rate and zero-miss false-alarm rate for that detector.

11.4 Results

Table 11.1 shows the average equal-error rates and their standard deviations for each of the 14 detectors over all 51 subjects from the lab and field data set respectively.

	Detector (lab data)	**equal-error rate**		**Detector (field data)**	**equal-error rate**
1	Manhattan (scaled)	0.096 (0.069)	1	Manhattan (scaled)	0.099 (0.095)
2	Nearest Neighbor (Mahalanobis)	0.100 (0.064)	2	Nearest Neighbor (Mahalanobis)	0.102 (0.106)
3	Outlier Count (z-score)	0.102 (0.077)	3	SVM (one-class)	0.106 (0.102)
4	SVM (one-class)	0.102 (0.065)	4	Outlier Count (z-score)	0.111 (0.108)
5	Mahalanobis	0.110 (0.065)	5	Mahalanobis (normed)	0.112 (0.103)
6	Mahalanobis (normed)	0.110 (0.065)	6	Mahalanobis	0.112 (0.103)
7	Manhattan (filter)	0.136 (0.083)	7	Manhattan (filter)	0.123 (0.077)
8	Manhattan	0.153 (0.092)	8	Manhattan	0.136 (0.085)
9	Neural Network (auto-assoc)	0.161 (0.080)	9	Neural Network (auto-assoc)	0.149 (0.080)
10	Euclidean	0.171 (0.095)	10	Euclidean	0.151 (0.081)
11	Euclidean (normed)	0.215 (0.119)	11	Euclidean (normed)	0.178 (0.099)
12	Fuzzy Logic	0.221 (0.105)	12	Fuzzy Logic	0.212 (0.142)
13	k Means	0.376 (0.158)	13	k Means	0.275 (0.131)
14	Neural Network (standard)	0.828 (0.148)	14	Neural Network (standard)	0.784 (0.152)

Table 11.1: The average equal-error rates for lab data (left side) and field data (right side) from the evaluation of the 14 detectors, ranked from best to worst (with standard deviations in parentheses).

Finding: The set of top-performing detectors is similar for lab and field data in terms of equal-error rate. We were surprised to find that the rankings are almost the same for both the lab and field data sets. The top four detectors when running on lab data are the same as those when running on field data. The only difference is that the Outlier Count (z-score) detector had greater equal-error rate than the SVM (one-class) detector when running on field data. The Manhattan (scaled) detector had the best performance running on both the lab and field data set.

Table 11.2 shows the performance difference of the 14 detectors, field vs. lab, in terms of equal-error rate.

	Detector	equal-error rate (Lab)	equal-error rate (Field)	Difference
1	Manhattan (scaled)	0.096	0.099	0.003
2	Nearest Neighbor (Mahalanobis)	0.100	0.102	0.002
3	Outlier Count (z-score)	0.102	0.111	0.009
4	SVM (one-class)	0.102	0.106	0.004
5	Mahalanobis	0.110	0.112	0.002
6	Mahalanobis (normed)	0.110	0.112	0.002
7	Manhattan (filter)	0.136	0.123	-0.013
8	Manhattan	0.153	0.136	-0.017
9	Neural Network (auto-assoc)	0.161	0.149	-0.012
10	Euclidean	0.171	0.151	-0.020
11	Euclidean (normed)	0.215	0.178	-0.037
12	Fuzzy Logic	0.221	0.212	-0.009
13	k Means	0.376	0.275	-0.101
14	Neural Network (standard)	0.828	0.784	-0.044

Table 11.2: Performance difference of the 14 detectors running on lab and field data in terms of equal-error rate. The 14 detectors are ranked from best to worst based on their performance on lab data. A positive difference indicates lab performance was better. A negative difference indicates field performance was better.

Finding: Performance difference between lab and field is small in terms of equal error rate. As shown in Table 11.2, for the top 6 detectors (rank 1 - 6), all of them had slightly worse performance when running on field data. Performance differed by less than 1% for the top 6 detectors, between lab and field data. For the rest of the 14 detectors, although the performance difference increased for some, their performance, even on lab data alone, was generally poor, and thus did not have a significant impact.

Table 11.3 shows the average zero-miss false-alarm rates for each of the 14 detectors, over all 51 subjects from the lab and field data sets, respectively.

	Detector (lab data)	**zero-miss false-alarm rate**		**Detector (field data)**	**zero-miss false-alarm rate**
1	Nearest Neighbor (Mahalanobis)	0.468 (0.272)	1	SVM (one-class)	0.460 (0.313)
2	Mahalanobis	0.482 (0.273)	2	Mahalanobis (normed)	0.464 (0.279)
3	Mahalanobis (normed)	0.482 (0.273)	3	Mahalanobis	0.464 (0.279)
4	SVM (one-class)	0.504 (0.316)	4	Nearest Neighbor (Mahalanobis)	0.467 (0.281)
5	Manhattan (scaled)	0.601 (0.337)	5	Manhattan (scaled)	0.487 (0.330)
6	Manhattan (filter)	0.757 (0.282)	6	Manhattan (filter)	0.645 (0.318)
7	Outlier Count (z-score)	0.782 (0.306)	7	Manhattan	0.722 (0.313)
8	Manhattan	0.843 (0.242)	8	Neural Network (auto-assoc)	0.747 (0.293)
9	Neural Network (auto-assoc)	0.859 (0.221)	9	Outlier Count (z-score)	0.756 (0.339)
10	Euclidean	0.875 (0.200)	10	Euclidean	0.759 (0.291)
11	Euclidean (normed)	0.911 (0.148)	11	Euclidean (normed)	0.782 (0.291)
12	Fuzzy Logic	0.935 (0.108)	12	Fuzzy Logic	0.860 (0.248)
13	k Means	0.987 (0.041)	13	k Means	0.963 (0.184)
14	Neural Network (standard)	1.000 (0.000)	14	Neural Network (standard)	1.000 (0.000)

Table 11.3: The average zero-miss false-alarm rates for lab data (left side) and field data (right side) from the evaluation of the 14 detectors are ranked from best to worst (with standard deviations in parentheses).

Finding: Performance results were similar between lab and field data sets in terms of zero-miss false-alarm rate. Based on the ranking shown in Table 11.3, the top 4 detectors were the same for both the lab and field data. However, the order for the top 4 detectors were reversed in the field data. The Nearest Neighbor (Mahalanobis) detector was the best-performing detector in lab data with zero-miss false-alarm rate of 0.468. The SVM (one-class) detector was the best-performing detector in field data with zero-miss false-alarm rate of 0.460. The five worst-ranked detectors (rank 10 - 14) were the same for both lab and field data, and they were in the same order as well. The zero-miss false-alarm rates were not very good for either data set. With the best detector (SVM) being 46%, these slight differences did not have a substantial impact on overall performance.

Table 11.4 shows the performance difference of the 14 detectors in terms of zero-miss false-alarm rate. For all detectors, their performance was slightly better when running on field data except for one detector. The Neural Network (standard) detector showed no improvement.

	Detector	zero-miss false-alarm rate (Lab)	zero-miss false-alarm rate (Field)	Difference
1	Nearest Neighbor (Mahalanobis)	0.468	0.467	-0.001
2	Mahalanobis	0.482	0.464	-0.018
3	Mahalanobis (normed)	0.482	0.464	-0.018
4	SVM (one-class)	0.504	0.460	-0.044
5	Manhattan (scaled)	0.601	0.487	-0.114
6	Manhattan (filter)	0.757	0.645	-0.112
7	Outlier Count (z-score)	0.782	0.756	-0.026
8	Manhattan	0.843	0.722	-0.121
9	Neural Network (auto-assoc)	0.859	0.747	-0.112
10	Euclidean	0.875	0.759	-0.116
11	Euclidean (normed)	0.911	0.782	-0.129
12	Fuzzy Logic	0.935	0.860	-0.075
13	k Means	0.987	0.963	-0.024
14	Neural Network (standard)	1.000	1.000	0.000

Table 11.4: Performance difference of the 14 detectors running on lab and field data in terms of zero-miss false-alarm rate. The 14 detectors are ranked from best to worst based on their performance on lab data. Positive difference indicates lab performance was better. Negative difference indicates field performance was better.

11.5 Discussion

Despite the observed differences between lab and field data, there is little performance difference between lab and field data sets in terms of equal-error rate and zero-miss false-alarm rate. The uncontrolled field environment little affected detector performance.

11.6 Summary

In this section, we answer the question: what is the difference in anomaly detector performance when operating on lab data versus field data? The findings can be summarized as follows:

- The set of top-performing anomaly detectors were the same when running on lab and field data in terms of equal-error rate and zero-miss false-alarm rate.
- Differences in performance of the 14 detectors were small.

Chapter 12

Training & Testing - Practice Effect

12.1 Research Question

Does practice influence impostor-detector performance differently between field and lab data?

12.2 Motivation

We have seen that practice can have a large impact on keystroke data. It stands to reason that it may also impact the performance of keystroke dynamics detectors. Knowing the difference in the practice effect on detector performance between lab and field data could influence detector training and testing procedures.

As we have identified in Part III, field subjects became practiced faster than lab subjects on average. The average field subject became practiced during Session 4 (repetition 197). If detector performance becomes stable after subjects become practiced, then we can train the detector on Session 5 data and test it on Session 6 data, instead of using Session 7 and Session 8, which is what we did for our study. This can help us reduce data collection time and research expenses.

12.3 Method

We used the Manhattan (scaled) detector in this study, because we had previously identified it as the top-performing detector. In this study, the data used to train and test the detector is on a session basis (50 repetitions). We started by selecting one session as the training data, and a different session as the testing data. We selected a genuine user; the rest of the 51 subjects were treated as impostors. During the training phase, we used the 50 repetitions in the training session by the genuine user to train the detector. During the test phase, we ran the detector on the 50 repetitions in the testing session by the genuine user to generate user scores. After that, we ran the detector on the first five repetitions from each impostor to generate impostor scores. We then repeated this process by assigning each subject as the genuine user. An equal-error rate was computed for each subject, which we averaged to form a system equal-error rate. We then repeated this procedure for each pair of sessions. We did not train and test the detector using the same session, since user scores generated that way are not meaningful.

12.4 Results

Table 12.1 comes from Lau and Maxion [2015] and shows the equal-error rate of the detector when using different training and testing sessions from lab data. Table 12.2 shows the equal-error rate of the detector when using different training and testing sessions from field data.

Comparing Table 12.1 and Table 12.2, we observe similar patterns. First, in both data sets, as the training session becomes farther apart from the test session, the EER increases. For example, when the detector is trained on field data from Session 1, the EER gradually increases as the test data changes from Session 2 to Session 8. Similarly, if the detector is tested on data from Session 8, the EER gradually decreases as the training data changes from Session 1 to Session 7. These two scenarios correspond to the first row and last column of Table 12.1 and Table 12.2.

	Testing 1	Testing 2	Testing 3	Testing 4	Testing 5	Testing 6	Testing 7	Testing 8
Training 1		11.8	11.2	12.8	13.4	14.2	15.6	16.8
Training 2	14.7		9.3	8.7	9.0	8.9	10.1	10.9
Training 3	16.2	10.8		7.9	8.4	8.6	9.4	9.6
Training 4	18.9	11.6	9.6		7.4	7.2	8.1	7.8
Training 5	19.1	12.3	10.5	7.8		6.7	7.3	8.0
Training 6	20.7	12.2	10.8	7.9	6.9		6.7	7.1
Training 7	22.1	15.0	12.1	9.0	7.9	7.5		6.9
Training 8	22.3	14.0	11.7	8.7	8.4	6.9	6.1	

Table 12.1: **Equal Error Rates (EER)** for Lab Data. The table shows the Equal Error Rate (EER) when using different training and testing sessions. The diagonal is left blank on purpose since we do not train and test on the same session.

Secondly, in both data sets, we notice a gradual decrease in EER in the upper diagonal of Table 12.1 and Table 12.2. As subjects get more and more practice on typing the password, the data becomes less noisy and thus increases the accuracy of the detector.

However, the detector performance did not become stable after field subjects became practiced. In Part III, we showed that the average field subject became practiced during Session 4 (repetition 197). Hence, data from Sessions 5 and 6 are practiced. When we trained the detector on Session 5 data and tested it on Session 6 data, it did not show similar result when trained and tested on Session 7 and 8. In fact, the equal-error rate was 7.9$\%$ when training and testing on Sessions 5 and 6, while the equal-error rate was 6.7$\%$ when training and testing on Sessions 7 and 8.

	Testing 1	Testing 2	Testing 3	Testing 4	Testing 5	Testing 6	Testing 7	Testing 8
Training 1		10.8	10.5	11.6	12.0	14.0	14.2	15.4
Training 2	15.1		7.6	9.0	8.5	10.1	10.5	10.5
Training 3	16.9	10.8		8.3	7.9	8.8	9.2	9.8
Training 4	19.0	12.3	8.7		7.2	8.3	8.1	8.3
Training 5	19.3	12.5	9.2	8.3		7.9	8.1	8.4
Training 6	21.7	14.6	10.8	9.8	8.7		6.7	7.3
Training 7	21.0	14.5	11.1	10.4	9.1	7.1		6.7
Training 8	22.3	14.8	11.4	10.3	9.2	7.4	6.2	

Table 12.2: **Equal Error Rates (EER)** for Field Data. The table shows the Equal Error Rate (EER) when using different training and testing sessions. The diagonal is left blank on purpose since we do not train and test on the same session.

12.5 Discussion

We observed in both data sets that detector performance generally became better as subjects became more and more practiced. Practice had similar impacts on detector performance for both data sets.

We were surprised that the detector performance did not become stable after subjects became practiced. Since field subjects became practiced more quickly than lab subjects did, we thought training on earlier sessions could have similar performance results as compared to training on Session 7. There might be internal variability, in addition to a practice effect, that was affecting detector performance. However, we are not sure what that is at this moment. Since the average lab subject became practiced during Session 6 (repetition 292), we were unable to further investigate this observation on the lab data set.

12.6 Summary

This section answers the question: does practice influence impostor-detector performance differently between field and lab data? The findings can be summarized as follows:

- We observed similar changes in equal-error rate for lab and field data. When the detector is trained and tested on more-practiced data, its performance improves.
- We did not observe stable detector performance after field subjects became practiced.

From the findings, we can see that field and lab data sets are quite similar in terms of practice effect on detector performance.

Chapter 13

Conclusion

In this study, we collected a field-based dataset that matches an existing lab dataset, each containing 51 subjects completing 400 repetitions of the password ".tie5Roanl". We used a newly-developed remote data-collection system to achieve this. Through the analysis on demographic information surrounding the two datasets, we discovered that the populations in both studies were similar in terms of gender and handedness, although the field subjects were more concentrated in the 21-25 age group. With basic descriptive statistics tools, we saw that the typing data in the two datasets became more similar to one other as subjects got more practice; that is, unpracticed subjects showed differences between the datasets, whereas highly practiced subjects did not. After analyzing the underlying distributions, we found significant differences between the two datasets, but these differences showed a very small effect size, rendering the two datasets essentially the same. By exploring the practice effect on typing data, we observed that field subjects became practiced more quickly (in fewer repetitions) and total typing time became similar for both datasets after practice. Despite minor differences between the two datasets, these differences had negligible effects on detector performance on lab vs. field data.

We conclude that lab and field data are interchangeable. From our own experience, field data can be less costly to collect since there is no involvement of research personnel. For our previous lab experiment, the equipment required for the study cost 3000 dollars. For each typing session, a subject needed to reserve half an hour (4 hours total for each subject) to complete the entire experiment. Our research assistant had an hourly rate of 45 dollars and thus cost 9180 dollars to finish the data collection of 51 subjects. Moreover, we needed to compensate subjects for their time. The lab experiment is 3000 + 9180 = 12180 dollars more expensive than the field experiment, considering that subject compensations are the same for both studies. Moreover, the field subjects can complete a typing session outside of the 9am-5pm time frame, and hence have more freedom in choosing their own schedule. Additionally, our field subject pool is not limited to the local population. Some of the subjects completed the experiment in Britain from a different time zone. Considering that practice can be achieved more quickly in the field, future researchers can collect more-practiced data at lower cost by using field data instead of lab data.

Appendix A

Five-number summaries of feature data

Table A.1 and Table A.2 show the five-number summary plus mean for each hold-time feature from Session 1 and Session 8 of lab data respectively. The units for values in the tables are seconds.

Table A.3 and Table A.4 show the five-number summary plus mean for each digram-latency-time feature from Session 1 and Session 8 of lab data respectively. The units for the values in the tables are seconds.

Table A.5 and Table A.6 show the five-number summaries plus mean for individual hold-time features from Session 1 and Session 8 of field data respectively. The units for the values in the tables are seconds.

Table A.7 and Table A.8 show the five-number summaries and mean for individual digram-latency-time features from Session 1 and Session 8 from field data respectively. The units for the values in the tables are seconds.

Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
H.period	0.0188	0.0731	0.0892	0.0921	0.1082	0.2974
H.t	0.0172	0.0628	0.0768	0.0815	0.0942	0.2184
H.i	0.0087	0.0607	0.0761	0.0793	0.0954	0.2441
H.e	0.0140	0.0648	0.0799	0.0845	0.0990	0.2085
H.five	0.0058	0.0604	0.0726	0.0760	0.0882	0.1875
H.Shift.r	0.0122	0.0665	0.0861	0.0903	0.1124	0.2318
H.o	0.0103	0.0673	0.0816	0.0844	0.0982	0.2468
H.a	0.0090	0.0765	0.0964	0.1027	0.1198	0.7221
H.n	0.0114	0.0641	0.0799	0.0827	0.0966	0.1932
H.l	0.0095	0.0744	0.0897	0.0909	0.1056	0.1995
H.Return	0.0064	0.0705	0.0853	0.0882	0.1035	0.2428

Table A.1: Five-number summary for Session 1 hold time features from lab data. The units are seconds.

Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
H.period	0.0069	0.0741	0.0886	0.0953	0.1077	0.2943
H.t	0.0098	0.0678	0.0842	0.0891	0.1050	0.2211
H.i	0.0050	0.0621	0.0774	0.0847	0.1025	0.3312
H.e	0.0077	0.0712	0.0854	0.0950	0.1131	0.3154
H.five	0.0066	0.0604	0.0767	0.0781	0.0923	0.1989
H.Shift.r	0.0114	0.0704	0.0943	0.0984	0.1199	0.2390
H.o	0.0140	0.0718	0.0892	0.0914	0.1064	0.2387
H.a	0.0195	0.0855	0.1048	0.1089	0.1254	0.3471
H.n	0.0064	0.0700	0.0899	0.0930	0.1113	0.3452
H.l	0.0061	0.0800	0.0969	0.0981	0.1146	0.2487
H.Return	0.0058	0.0708	0.0889	0.0914	0.1060	0.2566

Table A.2: Five-number summary for Session 8 hold time features from lab data. The units are seconds.

Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
UD.period.t	-0.0376	0.0938	0.1734	0.2583	0.2987	5.4890
UD.t.i	-0.0472	0.0288	0.0628	0.1119	0.1110	3.9160
UD.i.e	-0.0866	0.0238	0.0575	0.1154	0.1130	25.9200
UD.e.five	0.0501	0.2194	0.3748	0.4674	0.6111	4.5750
UD.five.Shift.r	0.1056	0.2910	0.3744	0.4778	0.5180	8.2910
UD.Shift.r.o	-0.0644	0.0823	0.1430	0.2143	0.2464	2.7480
UD.o.a	-0.2045	0.0257	0.0566	0.0893	0.0994	1.9590
UD.a.n	-0.2355	0.0138	0.0457	0.0796	0.1019	2.5240
UD.n.l	-0.0848	0.0749	0.1217	0.1746	0.2057	3.9780
UD.l.Return	-0.0171	0.1410	0.2069	0.3155	0.3360	5.8360

Table A.3: Five-number summary for Session 1 digram-latency-time features from lab data. The units are in seconds.

Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
UD.period.t	-0.1770	0.0332	0.0746	0.1293	0.1744	1.7720
UD.t.i	-0.1441	0.0235	0.0532	0.0701	0.0884	0.9263
UD.i.e	-0.1555	-0.0029	0.0299	0.0531	0.0761	1.5090
UD.e.five	-0.1257	0.1063	0.1503	0.2019	0.2436	2.5040
UD.five.Shift.r	0.0856	0.2006	0.2680	0.3180	0.3597	5.2010
UD.Shift.r.o	-0.0803	0.0388	0.0826	0.1310	0.1634	1.3270
UD.o.a	-0.1131	0.0107	0.0370	0.0556	0.0702	1.1610
UD.a.n	-0.1856	-0.0195	0.0103	0.0259	0.0499	2.0780
UD.n.l	-0.1201	-0.0024	0.0777	0.0870	0.1308	1.7990
UD.l.Return	-0.1245	0.1024	0.1439	0.1913	0.2233	2.2200

Table A.4: Five-number summary for Session 8 digram-latency-time features from lab data. The units are in seconds.

Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
H.period	0.0160	0.0789	0.0924	0.0980	0.1135	0.2400
H.t	0.0319	0.0785	0.0886	0.0931	0.1080	0.2310
H.i	0.0157	0.0680	0.0807	0.0893	0.1071	0.3525
H.e	0.0318	0.0773	0.0936	0.0964	0.1120	0.4142
H.five	0.0010	0.0721	0.0880	0.0901	0.1040	1.9200
H.Shift.r	0.0023	0.0800	0.1004	0.1018	0.1200	0.2332
H.o	0.0013	0.0720	0.0880	0.0951	0.1120	0.2400
H.a	0.0015	0.0875	0.1031	0.1073	0.1219	0.2559
H.n	0.0014	0.0720	0.0896	0.0975	0.1164	0.2241
H.l	0.0015	0.0797	0.0950	0.0989	0.1184	0.2177
H.Return	0.0047	0.0704	0.0864	0.0887	0.1040	0.2295

Table A.5: Five-number summary for Session 1 hold time features from field data. The units are in seconds.

Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
H.period	0.0042	0.0788	0.0958	0.0972	0.1132	2.0200
H.t	0.0052	0.0800	0.0960	0.0974	0.1120	1.9170
H.i	0.0035	0.0714	0.0879	0.0912	0.1075	0.9173
H.e	0.0027	0.0794	0.0960	0.0971	0.1137	0.2160
H.five	0.0015	0.0720	0.0877	0.0871	0.1026	0.4023
H.Shift.r	0.0015	0.0823	0.1044	0.1052	0.1279	0.2194
H.o	0.0016	0.0734	0.0919	0.0962	0.1140	1.5900
H.a	0.0029	0.0880	0.1059	0.1084	0.1279	3.6780
H.n	0.0032	0.0800	0.0963	0.1035	0.1220	0.7261
H.l	0.0027	0.0800	0.0977	0.1007	0.1200	0.2952
H.Return	0.0063	0.0718	0.0879	0.0978	0.1068	8.6880

Table A.6: Five-number summary for Session 8 hold time features from field data. The units are in seconds.

Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
UD.period.t	-0.0640	0.0687	0.1472	0.2228	0.3037	8.9200
UD.t.i	-0.1510	0.0105	0.0400	0.0654	0.0800	1.5760
UD.i.e	-0.1040	0.0110	0.0425	0.0698	0.0848	1.3920
UD.e.five	-0.0585	0.1512	0.2760	0.3911	0.5120	3.1920
UD.five.Shift.r	0.0018	0.2402	0.3426	0.4107	0.4800	10.1800
UD.Shift.r.o	-0.0483	0.0560	0.1001	0.1662	0.1908	19.7900
UD.o.a	-0.1840	0.0164	0.0509	0.0812	0.0985	1.8760
UD.a.n	-0.1391	-0.0123	0.0319	0.0576	0.0887	1.3520
UD.n.l	-0.1280	-0.0185	0.0876	0.0976	0.1524	1.7680
UD.l.Return	-0.0753	0.1200	0.1836	0.2833	0.3039	30.2500

Table A.7: Five number summary for Session 1 digram-latency-time features from field data. The units are in seconds.

Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
UD.period.t	-1.9160	0.0240	0.0703	0.1244	0.1486	2.4400
UD.t.i	-0.9146	0.0013	0.0319	0.0473	0.0651	0.9281
UD.i.e	-0.1372	-0.0115	0.0240	0.0409	0.0640	1.6760
UD.e.five	-0.0801	0.0960	0.1428	0.1944	0.2304	1.9600
UD.five.Shift.r	0.0023	0.1749	0.2450	0.2842	0.3316	3.5520
UD.Shift.r.o	-0.1040	0.0320	0.0752	0.1225	0.1586	17.6300
UD.o.a	-0.2080	0.0080	0.0398	0.0476	0.0721	0.9435
UD.a.n	-0.1520	-0.0343	0.0002	0.0106	0.0468	0.9601
UD.n.l	-0.2239	-0.0399	0.0391	0.0561	0.1199	1.7490
UD.l.Return	-0.1995	0.0779	0.1361	0.1576	0.2000	1.9120

Table A.8: Five number summary for Session 8 digram-latency-time features from field data. The units are in seconds.

Bibliography

Age and sex composition: 2010. May 2011. URL `http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf`. 5.5

William Lowe Bryan and Noble Harter. Studies in the physiology and psychology of the telegraphic language. *Psychological Review*, 4(1):27–53, January 1897. 2

J. Cohen. A power primer. *Psychological Bulletin*, 112:155–159, 1992. 7.3

Lawrence T. DeCarlo. On the meaning and use of kurtosis. *Psychological Methods*, 2(3): 292–307, 1997. 7.3

R. S. Gaines, William Lisowski, S. J. Press, and Norman Shapiro. Authentication by keystroke timing: Some preliminary results. Technical Report RAND-R-2526-NSF, Rand Corporation, Santa Monica, CA, May 1980. 2

Daniele Gunetti and Claudia Picardi. Keystroke analysis as a tool for intrusion detection. In Ahmed Awad E. Ahmed and Issa Traore, editors, *Continuous Authentication Using Biometrics: Data, Models, Metrics*, pages 193–211. Information Science Reference, Hershey, PA, 2012. 2

Ken Kelley and Kristopher J. Preacher. On effect size. *Psychological Methods*, 17(2): 137–152, 2012. 7.3

K. Killourhy and R. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. *IEEE/IFIP International Conference on Depentable Systems & Networks*, pages 125–134, 2009a. 11.3

Kevin S. Killourhy and Roy A. Maxion. 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, 29 June - 02 July 2009b. IEEE Computer Society Press. Estoril, Lisbon, Portugal. 2

Shing-hon Lau and Roy Maxion. The effect of practice in keystroke dynamics. *ACM Transactions on Information and System Security*, 2015. In submission. 8.3, 9.4, 9.4, 12.4

NIST. NIST/SEMATECH e-Handbook of statistical methods. `http://www.itl.nist.gov/div898/handbook/`, April 2012. Accessed: 2015-06-15. 7.3

Mohammad S. Obaidat. A verification methodology for computer systems users. In *ACM Symposium on Applied Computing (SAC)*, pages 258–262, New York, NY, USA, 1995. ACM Press. 2

Alen E. Peacock, Xian Ke, and Matthew Wilkerson. Typing patterns: A key to user identification. *IEEE Security and Privacy*, 2(5):40–47, September/October 2004. 2

Toshiharu Samura and Haruhiko Nishimura. Keystroke timing analysis for individual identification in Japanese free text typing. In *ICCAS-SICE 2009*, pages 3166–3170, Tokyo, Japan, 18-21 August 2009. SICE. 2

Pin Shen Teh, Andrew Beng Jin Teoh, and Shigang Yue. A survey of keystroke dynamics biometrics. *The Scientific World Journal*, 2013, August 2013. doi: 10.1155/2013/408280. 2

M.B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika(Biometrika Trust)*, 55(1):1–17, 1968. 7.3

Enzhe Yu and Sungzoon Cho. Novelty detection approach for keystroke dynamics identity verification. In Jiming Liu, Yiu-ming Cheung, and Hujun Yin, 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, 21-23 March 2003. Springer Verlag. 2