UNIVERSITY OF CALIFORNIA, IRVINE

# Usability of Security Critical Protocols Under Adversarial Noise

DISSERTATION

submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Tyler Michael Kaczmarek

Dissertation Committee:
Professor Gene Tsudik, Chair
Professor Alfred Kobsa
Professor Bruce Berg

2018

# Dedication

To Oliver
I cannot wait to meet you.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

There is only space for one name on the title page of this dissertation, but it does not in any way reflect the truth of the matter. I would never have made it to the end of the marathon without the support, love, and patience of so many people.

First, and foremost, I have to thank my adviser, Gene Tsudik. Without him I would not know the first thing about conducting proper, rigorous research, and my prose would be... unduly flowery, at best. He has stood by and provided guidance through thick and thin, and I can not express my gratitude deeply enough.

I would also like to thank my defense committee, Gene Tsudik, Alfred Kobsa and Bruce Berg. Not only did they encourage me throughout the process of creating my dissertation, they were all close collaborators and mentors throughout my candidacy.

I am extremely thankful for my parents, brothers, and especially my wife, Tara Kaczmarek, who provided unconditional love and support even on the most frantic of all-nighters.

Finally, I would like to thank all of my labmates and collaborators in the SPROUT group at UC Irvine: Karim El Defrawy, Robert Sy, Sky Faber, Cesar Ghali, Ekin Oguz, Christopher Wood, Luca Ferretti, Tatiana Bradley, Norrathep Rattanavipanon, Xavier Carpent, Ercan Ozturk, Ivan Nunes, Yoshimichi Nakatsuka, Pier Paolo Tricomi and Andrew Paverd.

This dissertation is the result of several years of research conducted at UC Irvine. During that time, I was supported by the University of California's Dean's Fellowship, The Butterworth Fellowship, and NSF Grant CNS-1544373 "EAGER: Unattended/Automated Studies of Effects of Auditory Distractions on Users Performing Security-Critical Tasks".

# Curriculum Vitae

Tyler Michael Kaczmarek

EDUCATION

**Doctor of Philosophy in Computer Science** **2018**
University of California, Irvine *Irvine, California*

**Master of Science in Computer Science** **2015**
University of California, Irvine *Irvine, California*

**Bachelor of Science in Computer Science** **2013**
The George Washington University *Washington, D.C.*

RESEARCH EXPERIENCE

**Graduate Research Assistant** **2013-2018**
University of California, Irvine *Irvine, California*

**Summer Research Assistant** **2017**
MIT Lincoln Laboratory *Lexington, Massachusetts*
**Summer Research Assistant** **2014**
Hughes Research Laboratory *Malibu, California*

TEACHING EXPERIENCE

**Reader for Computer & Network Security (ICS 134)** **Winter 2017**
University of California, Irvine *Irvine, California*

**TA for Computer & Network Security (ICS 134)** **Winter 2016**
University of California, Irvine *Irvine, California*

**TA for Intermediate Programming (ICS33)** **Spring 2015**
University of California, Irvine *Irvine, California*

PAPERS IN SUBMISSION OR UNDER REVIEW

”Thermanator: Thermal Residue-Based Post Factum Attacks On Keyboard Password Entry.” European Symposium on Security and Privacy 2019

”Exploring Effects of Auditory Stimuli on CAPTCHA Performance.” European Symposium on Security and Privacy 2019

REFEREED MAGAZINE PUBLICATIONS

”An exploration of the effects of sensory stimuli on the completion of security tasks.” IEEE Security & Privacy 15.6 (2017): 52-60.

REFEREED CONFERENCE PUBLICATIONS

”Assentication: User De-authentication and Lunchtime Attack Mitigation with Seated Posture Biometric.” International Conference on Applied Cryptography and Network Security. Springer, Cham, 2018.

”Lights, Camera, Action! Exploring Effects of Visual Distractions on Completion of Security Tasks.” International Conference on Applied Cryptography and Network Security. Springer, Cham, 2017.

”Proactively Secure Cloud-Enabled Storage.” Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on. IEEE, 2017.

”Byzantine Fault Tolerant Software-Defined Networking (SDN) Controllers.” Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual. Vol. 2. IEEE, 2016.

”An Unattended Study of Users Performing Security Critical Tasks Under Adversarial Noise” USEC 2015: 23014

”Dispute Resolution in Accessible Voting Systems: The Design and Use of Audiotegrity.” VOTE-ID 2013: 127-141

# Abstract

Usability of Security Critical Protocols Under Adversarial Noise

By

Tyler Michael Kaczmarek

Doctor of Philosophy in Computer Science

University of California, Irvine, 2018

Professor Gene Tsudik, Chair

An increasing number of security-critical tasks require human involvement. These tasks assume that the human is the weakest point in the security chain, and are explicitly designed to be as robust as possible while remaining human-usable. Failures in performing such tasks are typically blamed on human error. However, the human's sensory environment is usually not taken into consideration. The Internet of Things's emergence has created settings where a user's sensory inputs can be controlled remotely. To the best of our knowledge, there has been no prior work to evaluate the potential impact of malicious sensory input on human performance of security tasks.

In this dissertation, we evaluate usability of several security-critical tasks under

differing forms of adversarial noise. Specifically, we conduct a series of unattended experiments to evaluate the impacts on subject failure rate and task completion times when attempting Bluetooth Pairing, CAPTCHA entry, and short-authentication-string entry when exposed to crafted auditory and visual stimuli. We conclude that there is a rich space for both beneficial sensory stimulation, as well as a broad attack surface for adversaries that control a user's sensory environment. Additionally, we find that the impacts on task performance caused by unexpected sensory stimulation can be generalized according to the Brain Arousal Model.

# Chapter 1

# Introduction

This dissertation describes our exploration of impacts of sensory stimulation on performance of security-critical tasks. In particular, it details our experience conducting a series of user studies to evaluate subject performance completing three different security-critical tasks under a myriad of sensory stimuli. These studies were conducted in a series of unattended experiments that demonstrate the value of the as-of-yet unexplored unattended experimental paradigm.

The main contributions of this work are:

1. We conducted an extensive user study evaluating impacts of auditory stimuli on subjects performing security-critical tasks. We accomplished this by administering a range of naturally occurring, static sound stimuli, as well

as a single manufactured dynamic sound stimulus to unsuspecting subjects attempting Bluetooth Pairing. We found that the static stimuli have a strong positive effect on task completion rates, and that the dynamic stimulus had a strong negative effect.

2. We explored effects of visual sensory distractions on subject task performance. To this end, we experimented with a large number of subjects who were exposed to a range of unexpected visual stimuli while attempting to perform a security-critical task. Our results clearly demonstrate substantially increased task completion times and markedly lower task success rates.
3. We carried out a comprehensive user study on subjects responding to CAPTCHA challenges. Its results show that various types of auditory stimuli impact performance differently. While, as expected, highly dynamic stimuli degrade performance more than their static counterparts, the greatest impacts on subject performance was found for stimuli which were either biologically significant or task-specific.
4. Finally, we describe our exploration of the impacts of auditory stimuli on subjects performing timed short-authentication-string entry, a common task used as a second factor for authentication. We find that, while there are

some impacts on subject failure rates related to exposure to dynamic stimuli, cognitive simplicity of the task gives too much leeway to cause substantive negative effects.

The rest of this chapter covers: increasing ubiquity of Internet of Things (IoT) devices and emergence of the cyber-physical sensory environment, psychological relationship between sensory stimulation and general task performance, and security protocols developed with human involvement.

## 1.1 Emergence of Cyber-Physical Sensory Environments

The emergence of the Internet of Things represents the shift of traditional networking concerns from connectivity of general-purpose machines to highly specific, lightweight sensor networks[18]. IoT devices, especially in the context of smart homes, have exploded in popularity over the two decades since the term was coined [67]. It is estimated that, by the end of 2018, there will be over 1.2f billion smart home IoT devices [1]. Smart home devices include everything from smart light bulbs to smart speakers, appliances and security systems [42]. These devices can be controlled remotely through voice, remote control, and computer applica-

tions either on a local network or across the Internet. These devices are marketed to average (i.e., not generally tech-savvy) end-users, and little evaluation has been done of their security implications of their iniquitous development [59]. Relevant prior research focused almost exclusively on securing the devices themselves [53]

Newly instrumented IoT environments allow for unprecedented fine-grained control of users' sensory environment, as now the lights, sound, temperature and more are controlled automatically and remotely with a greater deal of precision than can be accomplished in the traditional analog case [65].

This creates new opportunities for adversaries. Where before the compromise of a lighting or sound system required physical presence, now an adversary targeting a smart home can conduct an attack remotely. The sensory attack space available to an adversary who compromised the suite of devices has not been as thoroughly explored as securing the devices themselves. In particular, we are interested in the implications of smart devices in the typical instrumented home that control the occupants' sensory environment, such as smart speakers, which control the soundscape, and smart light bulbs, which control the visual environment.

## 1.2 Sensory Stimulation

Sensory stimulation can produce a wide spectrum of effects. In general, humans are adept at remaining focused on critical tasks despite interference due to induced stimulation. It might even be the case that the urgency or potential danger conveyed by different stimuli can serve to sharpen focus, and improve general task performance.

Any capture of an individual's attention by an aversive stimulus is likely to be momentary, occurring primarily when the stimulus is first introduced. In cognitive science, attention is conceptualized as a limited resource. Probably for good reason, the greatest demand on attention is in response to a change in the environment. Once an assessment is made that a stimulus does not require a response, adaptation to the stimulus from a foreground target into a background context proceeds relatively rapidly as attention is redistributed to other demands. Although an aversive sound may remain aversive throughout its presentation, its capacity to disrupt performance on a complex task might rapidly fade after onset. In fact, the introduction of such a sound can serve to increase an individual's overall awareness of thier sorroundings after the fact due to the elastic nature of the attentional resource [45].

Synthetic stimuli can be designed to attract attention resources without neces-

sarily being aversive. For instance, a crescendoing sound could embody a context of constant change, essentially "tricking" the system into a state of sustained engagement by creating the sensation that something is approaching that may represent a threat, without the sound itself being overtly dangerous or threatening. In general, synthetically constructed stimuli can be more engaging and have a greater sustained effect than their natural counterparts.

## 1.3 Security-Critical Protocols With Human Interaction

It is widely believed that the human user is the weakest link in the security chain [17]. Nonetheless, human participation is unavoidable in many security protocols. In fact, security protocols involving human users have become more commonplace in recent years [68]. Reliance on human involvement has created a new set of requirements for protocol development centered around tradeoffs between usability and security. This has led to a variety of efforts seeking to develop the most effective methods [5]. Such protocols require extensive usability testing, since users are unlikely to perform well when faced with overly difficult or intricate tasks. Typically, security-related usability testing entails evaluating human per-

formance in a ”best-case” scenario. In other words, testing is usually conducted in sterile lab-like environments. While appropriate for minimizing confounding environmental factors and focusing on protocol features, this does not reflect the average protocol use-case in the real world. In particular, effects of unexpected distractions on users participating in security protocols are unknown.

### 1.3.1 Bluetooth Pairing

The first Security-critical task that we evaluate is Bluetooth Pairing. A more detailed examiniation of the specifics of Bluetooth Pairing protocols are described in detail in Chapters 2 and 3. We present a brief summary of their development below.

Bluetooth wireless technology was developed in 1998 with the intention of providing a short-range connectivity solution for personal, portable devices [9]. In order for devices to establish a connection to one another to enable Bluetooth communication with one another, they must undergo a process known as Bluetooth pairing. Bluetooth pairing is typically conducted by a single user which is the owner of both devices.

Secure device pairing has been extensively researched by experts in both security and usability. While initially pairing, the two devices have no prior knowl-

edge of one another, i.e., there is no prior security context. Also, they can not rely on either a Trusted Third Party (TTP) or a Public Key Infrastructure (PKI) to facilitate the protocol. This makes device pairing especially vulnerable to man-in-the-middle (MiTM) attacks. This prompted the design of numerous protocols requiring human involvement (integrity verification) over some out-of-band (OOB) channel.

Many protocols have been developed to preform Bluetooth pairing. Proposed techniques range from entering an authentication string generated by each device onto the other [61], to taking photographs of a barcode identifying the device to pair with [12], to physically holding the devices and shaking them together [57].Short-Authentication-String comparison is one of the most common techniques, and is found in general to be highly usable [38]. The security task at the core of this method of Bluetooth pairing involves the user comparing two 6-digit decimal numbers – one displayed by each device being paired – and pressing a single button confirming a match. This is a discrete and uniform activity well suited for the creation of a homogenous user experience for experimentation[6].

### 1.3.2 CAPTCHA Challenges

Next, we look at CAPTCHA challenges. Chapter 4 describes our efforts in full detail, but a summary of the goals in the development of CAPTCHA techniques is provided below.

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) are programs that generate and evaluate challenges that are easy for a human to solve, yet cannot be easily solved by software. CAPTCHAs have been used to prevent bot-based abuse of services for over a decade. As CAPTCHA challenges became more commonplace, design decisions intended to increase the resilience to botting led to increasingly difficult to use challenges. Because of this, efforts were made to standardize their security properties and most recent efforts in development have been invested into creating CAPTCHAs that are [13]:

1. Usable: humans are successful at least $90\%$ of the time.
2. Secure/robust: a state-of-the-art bot should not be successful more than $0.01\%$ of the time.
3. Scalable: challenges that are either automatically generated, or drawn from a body that is too large to hard-code responses for each possible challenge.

Based on these requirements CAPTCHA developers focused on text-based

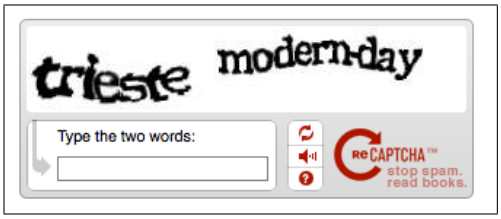


Figure 1.1: A Text-Based CAPTCHA.

CAPTCHAs, i.e., those that present a jumbled alphanumeric code as the challenge to the user, as shown in Figure 4.1. This approach is popular since human users are quite good at identifying these alphanumeric codes in an altered image, thus satisfying the usability requirement. Also, image segmentation and recovery is a known hard problem for AI, satisfying the security requirement. Finally, such challenges can be randomly generated as needed, thus satisfying the scalability requirement [21].

### 1.3.3 Two-Factor Authentication

Secure, correct and efficient user authentication is an integral component of any meaningful security system. Authentication schemes in the typical modern workplace typically include two factors. Many techniques have been proposed and evaluated. Authentication techniques fall into one of three types: 1.) What you know (e.g., a password,) 2.) What you own (e.g., token-based authentication,) or 3.) What you are (e.g., biometric authentication). Typically, the first factor in an

authentication is password/PIN entry, and falls into the ”what you know” category. While there are many interesting schemes based on biometric authentication, we did not evaluate their usability, as they are fraught with enrollment issues that would make effective large-scale studies in our unattended style infeasible [58].

We focus on a two-factor authentication flow where (1) the user demonstrates knowledge of a secret password or PIN, and (2) the user proves possession of a secure device or token [55]. This second factor seeks to avoid many of the problems associated with knowledge-based authentication by removing the burden of relying on a human to recall a complex string. Instead, it relies on using a secure hardware token or trusted smartphone application to generate a short-lived key that the user enters alongside their PIN or password [4]. This has led to relatively high adoption rates of smartphone applications such as DUO Mobile [24] and physical tokens such as the RSA Securid token [2].

# Chapter 2

# Effects of Auditory Noise on Completion of Security Critical Tasks

## 2.1 Introduction

Our world is a noisy and distracting place, where truly quiet or sterile environments are rare. Most people are accustomed to some degree of auditory and visual distraction in their daily lives. However, they may be influenced in an unexpected manner by sudden distractions, especially if they occur during performance of a

task that demands concentration.

Meanwhile, modern technology allows – and sometimes requires – people to engage in security-critical tasks in public settings, while being subjected to various degrees and types of sensory input. As personal wireless devices (mainly smartphones) become more ubiquitous, the average person grows more reliant on them for the performance of security tasks, such as entering a PIN, Bluetooth pairing or verifying transaction amounts. For example, in online fund transfers, one has to compare the displayed amount and currency to the intended amount and currency [23]. In device pairing, one needs to compare items (such as numbers, text, pictures, or sounds), or perform some physical task over an "out of band" (OOB) channel [35].

All these tasks require some form of human involvement, which represents the weakest link and determines overall security [23, 35, 47, 29, 30, 34, 51]. This motivates extensive usability studies to assess human ability to routinely complete security tasks that still provide an acceptable level of security. There has been a lot of research on this topic [47, 29], but very little work only that investigates user errors and maliciously induced user errors. One major reason for the dearth of prior work in this area is the difficulty of conducting traditional user experiments. Since human errors in such cases are relatively rare, it would take many trials with

many subjects to obtain statistically reliable information about the failure rate, and to determine whether the difference in rates between two methods is statistically significant. The problem is exacerbated by the fact that more than one method needs to be tested, while at the same time, only one attempt should be made per study participant to trigger a mistake (since subjects may otherwise become alerted to such attempts, consciously or subconsciously). For all these reasons, the total number of subjects needed for an experiment to study user mistakes in security-related tasks can quickly grow into the hundreds.

To mitigate the effort needed to conduct such large studies, we designed a setup for an entirely unattended experiment, wherein subjects receive recorded instructions from a life-size, video-projected, rather than ”live”, experimenter. As a first experiment in this environment, we decided to test the error rate of subjects attempting to pair two Bluetooth devices in the presence of unexpected audio stimuli. We tested $168$ subjects in this environment with no experimenter involvement. Our original expectation was that unexpected audio interference would have a negative impact on the completion of security-critical tasks. However, surprisingly, it turned out that noise actually had a facilitatory effect.

*Organization:* Section 5.2 presents the design and setup of our experiments. It is followed by Section 5.3 which presents experimental results. Next, Sections 3.5

and 5.4 summarize lessons learned from this experience and discuss conclusions, respectively. Section 3.6 acknowledges certain limitations of our approach. Then, Section 2.7 addresses ethical considerations and Section 5.8 concludes the chapter.

## 2.2 Experimental Setup and Methodology

This section describes our experimental setup, procedures and subject parameters.

Figure 2.1: Experimental Setup: (a) Side view (speakers over the door), (b) Front view

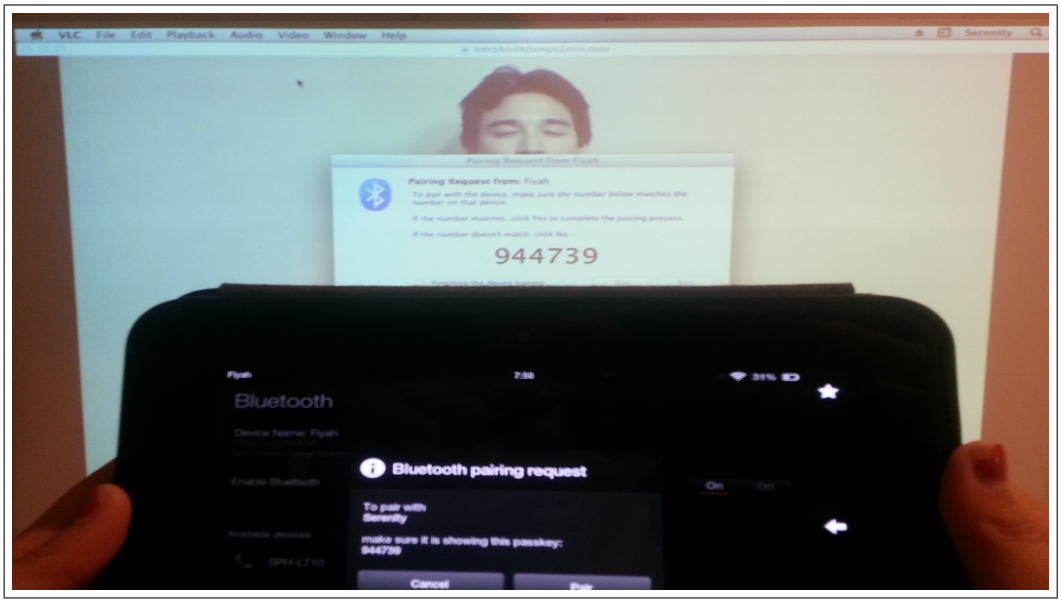


Figure 2.2: Bluetooth confirmation screen, from subject's perspective

### 2.2.1 Apparatus

The setting of our study was carefully designed to facilitate fully automated experiments with a variety of sensory inputs. The installation is situated in a low-traffic public space (a wide corridor corner nook) at the top floor of a large academic building on a university campus.

Figure 3.1(a) shows the experimental location from the side, and Figure 3.1(b) shows our setup from the subject's perspective (front view). It includes a large touch-sensitive Smartboard with a short-throw projector, a webcam, and two pairs of speakers (one in front and one behind the intended subject position), as well as controllable lights and electricity outlets. The Smartboard is an interactive white-

Figure 2.3: Experimenter proxy giving video instructions

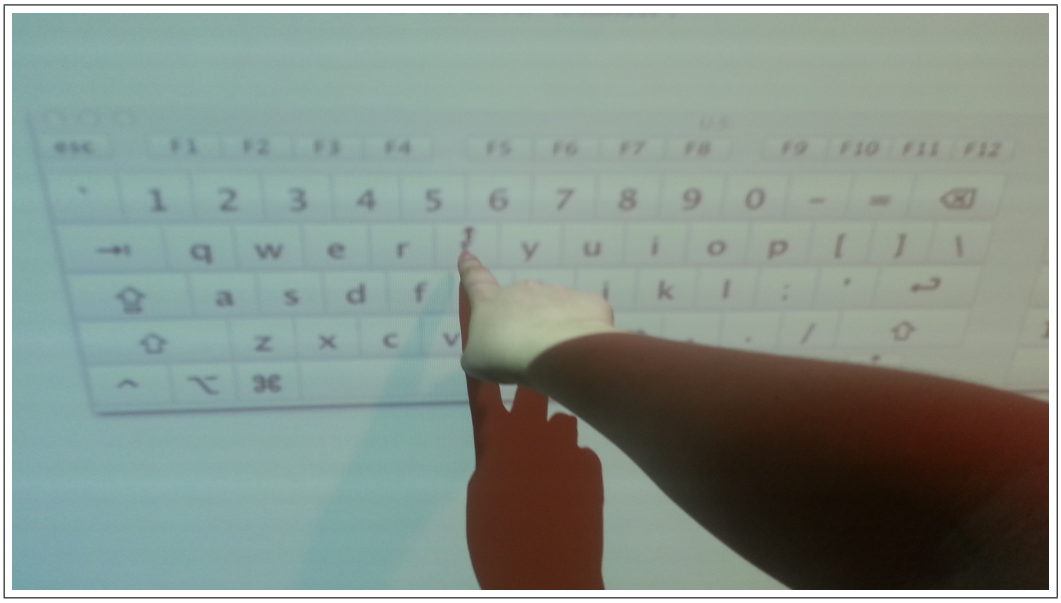

Figure 2.4: Subject entering email address on Smartboard

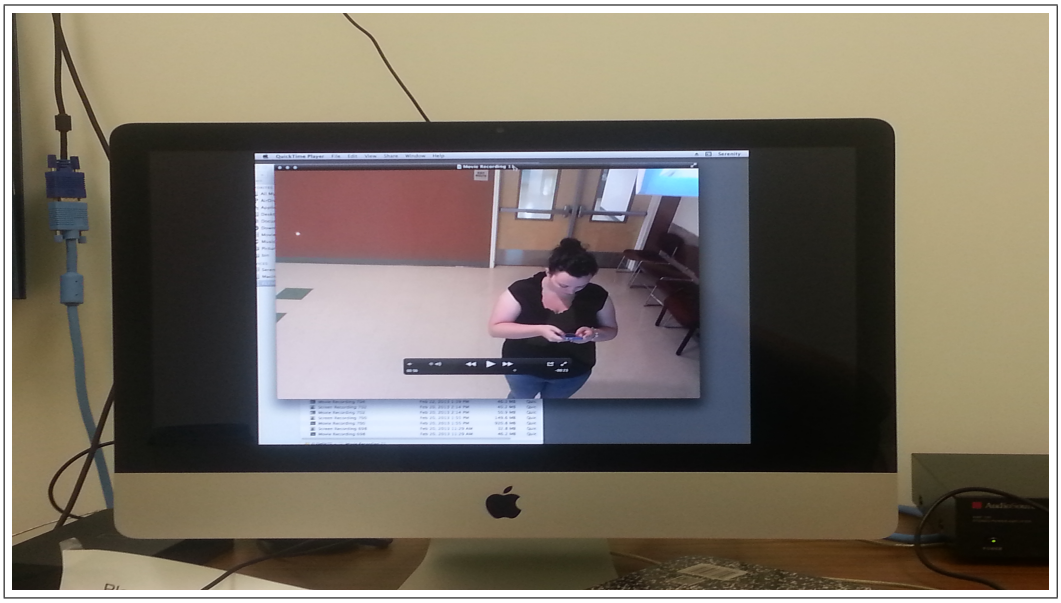

Figure 2.5: Post-experimental review of video recordings (separate office)

board (see `smarttech.com`) that gathers input via user's touch on its surface. As such, it acts as both the display and the input device.

Instead of a human experimenter actively curating the environment and interacting with the subjects, we used a life-size video/audio recording of an experimenter as a proxy. This proxy is the subject's main source of information about the experiment. In particular, the proxy starts by reading a script explaining the flow of the experiment. This is shown in Figure 2.3.

This setup allows for a fully unattended experiment. The only (and strictly off-line) involvement of a human experimenter amounted to the infrequent re-calibration of sound effect volume and repair of some components that suffered

(minor) damage throughout the study.

### 2.2.2 Procedures

The goal of our experiment was to measure user errors and average task duration while attempting to pair two wireless devices via Bluetooth, while being exposed to potentially distracting and possibly "malicious" auditory stimuli. We chose 5 different auditory stimuli, 4 seemingly innoccuous sounds that one encounters in real life, both in open and enclosed public spaces: (1) a baby crying, (2) a hammer stirking a wall, (3) heicopter rotors spinning and (4) a circular saw cutting wood as well as 1 manufactured dynamically "looming" sound described by 2.2.2. Reasons for selecting these four specific sounds as audio stimuli are discussed in Section 3.6.2 below.

The 4 natural sounds were played at a static volume from the speakers situatied behind the subject, while the dynamic looming sound was played from random speaker balances across all 4 speakers. Specific volumes of the five sounds (measured at a typical subject's position) were as follows:

- Baby: 67 dB
- Helicopter: 79 dB
- Hammer: 80 dB

- Saw: 78 dB

The volume of the dynamic looming stimulus increased from nearly silent to 85 dB over 5 seconds, its intensity curve is described in Figure 2.2.2. After the sound completed, it would repeat at a different Left/Right and Front/Back speaker balance, selected randomly. This would repeat continuously for the entirety of the two minute pairing window.

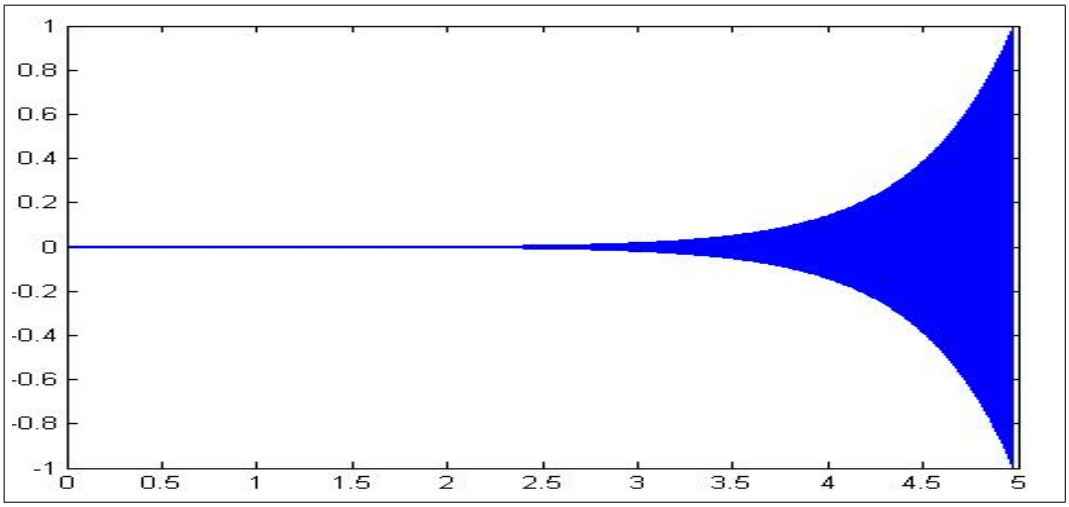


Figure 2.6: The Looming Sound Intensity Function

Even the highest of these five volumes (85 dB) is well within the *safe range*, as defined by the US Occupational Safety & Health Administration (OSHA) guidelines.[1]

[1]OSHA requires all employers to implement a Hearing Conservation Program where workers are exposed to a time-weighted average noise level of 85 dB or higher over an 8 hour work shift. Our noise levels were clearly lower and of a much shorter window of exposure. See: `https://www.osha.gov/SLTC/noisehearingconservation/`

To begin the experiment, the subject approaches the Smartboard and presses a large wall-mounted button to the right. Although a motion-activated start is also possible, we decided to minimize any disturbance for uninvolved passers-by. Next, the Smartboard plays a short video recording of our proxy experimenter, who explains that the subject will be preforming a task on their own phone, namely, connecting it via Bluetooth to a nearby device. The latter is actually an iMac desktop in the office behind the Smartboard; it is not visible to the subject, as shown in Figure 2.5. The subject is promised a reward for the successful completion of the experiment, in the form of a $\$5$ Amazon coupon. The subject is also briefly informed that the task of pairing two Bluetooth devices involves comparing two 6-digit numbers and confirming whether they match, as shown in Figure 2.2.

At this point, the subject has a time window of 2 minutes to correctly pair the devices. Otherwise, a failure message is read out and displayed. While the subject is in the process of pairing, one of five events occurs: either silence is maintained throughout the experiment, or one of the aforementioned four sounds is played from the speakers located on the ceiling behind the subject.

A subject who fails the first time and wishes to make another attempt at pairing, is given the opportunity to re-try the experiment in another two-minute win-

dow. If pairing completes successfully, a message to that effect is displayed. At the end, a subject is asked to enter an email address using a virtual keyboard displayed on the touch-sensitive Smartboard (see Figure 2.4), thus allowing us to email the promised Amazon coupon as a participation reward.

Each subject encountered only one condition. Presenting subjects with two conditions would have biased their performance in the second condition, since, at that point, they would already know what to do and what might happen. Since subject observables (errors) are influenced by various individual characteristics, random subject selection ensures that any variation between sample and population observables is only a matter of chance.

After successful completion of the experiment, if the same subject attempts to repeat the same experiment with the same personal device, their data is automatically flagged and later discarded. Multiple participation of the same subject with different personal devices is identified (and discarded) by visual inspection of video recordings. The experimental setup maintains a detailed log of all system events that can later be analyzed to measure outcomes, such as the number of retrials, task success rates, and task completion times, as well as a video recording of the entire encounter, as shown in Figure 2.5.

### 2.2.3 Hypotheses

Our initial hypotheses were that introducing noise while an unsuspecting subject attempts to pair two Bluetooth devices will have no effect:

H1 We will observe the same error rate

and

H2 The pairing process will take the same amount of time to complete successfully,

as in the same setting without any noise interference.

### 2.2.4 Subjects

In prior studies on usability of pairing protocols with human involvement [23], [47], [29], it was discovered that a subject population of 20-25 per condition being tested was an acceptable size for obtaining statistically significant findings. Since our planned experiment has one condition for each of the five sound effects as well as one control condition (with no sound), collecting any meaningful amount of data would require well over one hundred iterations of the experiment.

To recruit subjects, we posted signs around the entrance and inside the lobby of a large campus building, which directed people to the experimental setup and mentioned the reward for participation. Posters explicitly described that subjects

were sought for a brief ”Usability Study” and did not in any way mention the security-critical nature of the task to be performed, or the possibility of any noise interference. The general area of campus where the experiments were conducted houses Computer Science and Engineering departments.

Of the total 168 subjects, there were 115 males and 53 females. Most of them (159 out of 168) appeared to be college-aged (18-24 years), while 9 belonged to an older group (25+ years). This demographic breakdown is influenced by the location of the experiment and by the recruitment form. Since we solicited participants passively and since our recruitment posters were located in the “technical” part of a large university campus, it is not surprising that the overwhelming majority of participants were of college age with the majority being male.

## 2.3 Results

We now discuss the results of the study, starting with data cleaning and proceeding to task completion results.

### 2.3.1 Data Cleaning

Subject data was discarded in three cases. First, we removed the instances where participants arrived either in pairs or larger groups. Their data were eliminated since it might have been skewed due to social facilitation. It has been shown that being under observation of others can have a positive impact on subjects performing tasks of low levels of complexity [3]. Second, a few participants arrived with old-style flip phones. Such older phones were technically unable to establish a Bluetooth connection with our client.

All in all, 29 pairs or groups of subjects had to be discarded, as well as 10 others who attempted to use flip phones. We could not discern any obvious visual or auditory impairment in any subject that would be a detriment to the experiment. We later visually checked all experiments for subjects with such impairments and none were identified.

### 2.3.2 Task Completion Rate

Table 2.1 shows the numbers of subjects whose first attempt at pairing resulted in a success and failure, respectively, plus the failure rate for the control condition and each stimulus condition.

Table 3.2 shows the parameters for the Barnard's exact test applied pairwise

Table 2.1: Subject failure rate

| Stimulus | #Successful Subjects | #Unsuccessful Subjects | Failure Rate |
|---|---|---|---|
| None (control) | 27 | 13 | 0.34 |
| Baby | 23 | 1 | 0.04 |
| Hammering | 33 | 3 | 0.08 |
| Helicopter | 24 | 1 | 0.04 |
| Saw | 20 | 2 | 0.09 |
| Looming | 21 | 13 | 0.62 |
| **Total** | 148 | 33 | 0.22 |

to the subject failure rate of the control condition and each stimulus. It shows that differences between failure rates are statistically significant ($p < 0.05$) with respect to all four stimuli. This also holds if one applies a conservative Bonferroni correction to account for four pairwise comparison, which leads us to reject hypothesis H1 in Section 5.2.4, since the failure rate significantly decreases with the introduction of noise. Section 5.4 discusses this further.

Table 3.3 shows odds ratios and $95\%$ confidence interval for each stimulua compared to the control condition. Interestingly, under this analysis, confidence interval of the Saw condition includes a possible odds ratio of $1.0$. This implies that – under this method of analysis – it is not statistically significant at the $95\%$

Table 2.2: Barnard's Exact Test on subject failure rates of control & stimuli

| Stimulus | Total Pairings | Failure Rate | Wald Statistic | Nuisance Parameter | $p$ |
|---|---|---|---|---|---|
| None(control) | 40 | 0.34 | – | – | – |
| Baby | 24 | 0.04 | 2.65 | 0.95 | 0.03 |
| Hammering | 36 | 0.08 | 2.58 | 0.91 | 0.01 |
| Helicopter | 25 | 0.04 | 2.71 | 0.89 | 0.01 |
| Saw | 22 | 0.09 | 2.05 | 0.84 | 0.03 |
| Looming | 21 | 0.62 | 2.05 | 0.84 | 0.03 |

level. Confidence intervals for other 3 stimuli reinforce the claim of statistical significance at the $95\%$ level, as established by Barnard's exact test.

Table 2.3: Odds Ratio and $95\%$ Confidence Intervals on Subject Failure Rates of Control and Stimuli

| Stimulus | Odds Ratio wrt control | 95% Confidence Interval wrt control |
|---|---|---|
| None (control) | - | – |
| Baby | 0.09 | 0.01 - 0.74 |
| Hammering | 0.18 | 0.04 - 0.73 |
| Helicopter | 0.09 | 0.01 - 0.71 |
| Saw | 0.20 | 0.04 - 1.02 |

We also examined subject failure rates by gender, as shown by Tabe 3.4. While

it may appear at a cursory glance that female subjects were less likely to fail at Bluetooth pairing than thier male counterparts, performing Barnard's exact test on the subject failure rates of men and women revealed that the perceived difference between them is not statistically significant; Wald statistic = 0.64, nuisance parameter = 0.02, $p = 0.39$.

Table 2.4: Subject failure rate by gender

| Gender | #Successful Subjects | #Unsuccessful Subjects | Failure Rate |
|---|---|---|---|
| Male | 91 | 24 | 0.21 |
| Female | 45 | 9 | 0.17 |

### 2.3.3 Task Completion Times

Table 5.2 shows average completion times in successful trials for subjects under each stimulus. After applying a conservative Bonferroni correction to account for four pairwise comparisons, there is no statistically significant difference in completion times between the control condition and each stimulus.

Table 3.6 shows Cohen's $d$ and its 95% confidence interval, for subject completion times under each of the stimuli when compared to the control condition. The static sound stimuli do not show any statistically significant result, as thier

Table 2.5: Avg times (sec) for successful pairing

| Stimulus | Mean Time | Standard Deviation | DF wrt control | t-value wrt control | $p$ |
|---|---|---|---|---|---|
| None | 34.41 | 13.78 | – | – | – |
| Baby | 31.13 | 10.06 | 63 | 0.97 | 0.35 |
| Hammering | 28.82 | 9.76 | 74 | 1.84 | 0.07 |
| Helicopter | 31.33 | 13.13 | 63 | 0.81 | 0.39 |
| Saw | 38.45 | 17.15 | 60 | 0.90 | 0.38 |
| Looming | 80.75 | 10.40 | 38 | 10.03 | ¡ 0.01 |

confidence intervals contain 0. However, the dynamic looming stimulus does show a significant decrease in task completion speed at the $\alpha = 0.05$ level.

Table 2.6: Cohen's $d$ and 95% Confidence Intervals on Subject Completion Times Between Control and Stimuli

| Stimulus | Cohen's d wrt control | 95% Confidence Interval wrt control |
|---|---|---|
| None (control) | - | – |
| Baby | 0.27 | -4.00 to 4.29 |
| Hammering | 0.47 | -3.80 - 3.66 |
| Helicopter | 0.23 | -4.04 - 5.48 |
| Saw | -0.27 | -4.54 - 6.89 |
| Looming | -4.13 | –0.90 |

As with subject failure rates, we also examined subjects' completion times for

successful pairing attempts by gender. The results are displayed in Table VIII. A pairwise t-test shows that the observed differences are not statistically significant ($t(156) = 1.23$, $p = 0.22$).

Table 2.7: Avg times (sec) for successful pairing by gender

| Gender | Mean Time | Standard Deviation |
|---|---|---|
| Male | 30.63 | 10.92 |
| Female | 33.23 | 13.85 |

## 2.4 Lessons Learned

As mentioned above, some subjects participated in the experiment in pairs. We had not explicitly forbidden this since doing so in an unattended setting would be impossible. We ignored the data of such participant pairs, see Section 3.3.1. A few subjects also tried the experiment more than once on different Bluetooth devices (presumably to earn the participation reward multiple times), and we had to visually identify and discard their data.

Furthermore, a few subjects did not understand how to pair two devices using Bluetooth, or were unsure what they were supposed to do in general. This illustrates one drawback with our experiment design - there was no option to replay

the instructions, nor was there a set of more detailed instructions for participants who were unfamiliar with the Bluetooth functionality of their devices. Since our experiment was unattended, there was no way to tell the cause of task failure in real time, or to help the subject if needed, until the recording of the subject's trial was viewed.

Interestingly, quite a few subjects had trouble following the instructions of the proxy to enter their email address on a virtual keyboard that was projected onto the touch-sensitive Smartboard. Up to this point, the Smartboard had only served as a (completely passive) projection wall, and subjects may have been surprised that it could also be used as an input device.

## 2.5 Discussion of Observed Effects

The impact of unexpected peripheral auditory sitmuli to introduced to subjects performing security critical tasks differed by the type of stimulus used. The static, naturally occuring stimuli appeared to have a significant positive effect on subject success rates and an insignificant effect on subject completion speeds. Conversely, the dynamic, manufactured looming stimulus appeared to have a significant negative effect on both subject success rates and completeion speeds for successful

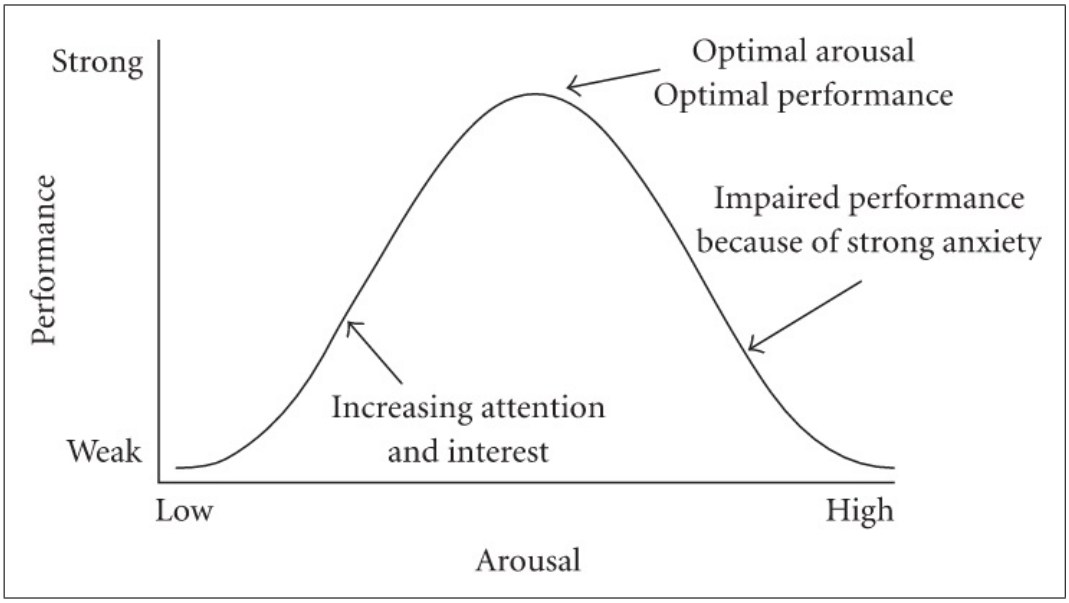


Figure 2.7: The Yerkes-Dodson Relationship Between Sensory Arousal Levels and Performance

trials. These results are consistent with the Yerkes-Dodson Law. This law states that subject task performance is related to thier overall attentional arousal level so that subjects at a very low or very high level of overall attentional arousal are likley to perform poorly on tasks, and subjects at a morderate level of attentional arousal are likely to perform optimally on tasks, as exemplified by Figure 5.2.

## 2.6 Limitations

We readily acknowledge that the study described in this chapter, although the first of its kind, has certain shortcomings and limitations, detailed below.

### 2.6.1 Subjects

We experimented with a narrow subject group, dominated by young and tech-savvy college students. This is a direct consequence of the specific campus location of our unattended setup. Replicating it in a non-academic setting (e.g., an office building) would be possible and useful. However, passive recruitment of a really diverse group of participants is only possible in a truly public space with a high volume of traffic, e.g., a stadium, a shopping mall, a movie theater or a concert hall. On the other hand, placing our unattended experiment setup in any of such settings would be extremely challenging. First, our setup involved specialized and expensive equipment whose security would be difficult to ensure in a very public space. Second, high-traffic public spaces tend to be have lots of background or ambient noise which would interfere with the stimuli in our experiments.

As already mentioned, the nature of our location also had a skewed impact on the gender breakdown of our subjects. Since the experiment was set up in the Computer Science and Engineering section of a large university campus, the majority of the passers-by were male. Because of this, we were unable to collect sufficient data in a realistic time frame to examine the effects of each individual stimulus on subjects of each gender.

One potential problem with our subjects is that young people are in general more sensitive to noise than older adults [10]. It is quite conceivable that older and/or technologically non-adept people[2] would react differently to our noise stimuli.

Finally, recall that our experimental setup required the subject to interact with both visual and audio queues from the proxy experimenter and the environment. Because of this, an ideal subject would have no substantial hearing or visual impairment. However, due to the unattended nature of our experiment, we could not proactively rule out such subjects (e.g., by specifying restrictions in the recruitment posters) without giving away the nature of our experiment. Doing so would have created an initial expectation for subjects who fit our criteria, which could adversely influence accuracy of collected data. Therefore, during later review of each video-recorded experiment, we had to verify that there were no participants with obvious visual and/or hearing impairments.

### 2.6.2 Diversity of Stimuli

We experimented with four stimuli through a subjective process of elimination, with the intention of getting as many diverse noise types as we could rigorously

[2]Since people who are new to, or unfamiliar with, a specific technological task would naturally be more nervous or tense when performing it.

test, that were annoying to the listener in varying degrees. With respect to diversity, we classified sounds in three ways:

1. Continuous or discrete
2. Regular or irregular
3. Human-generated or synthetic/mechanical

One reason for settling on such a small number of stimuli was due to the combination of (1) the location of the experiment, and (2) placement of study recruitment posters. Although posters were placed in a high-traffic zone, outside the building where the experiments took place, the same people (mostly students) tend to walk by every day due to the regularity of campus life, e.g., classes begin and end at the same time and at the same place. Consequently, although we were able to attract 147 subjects, the rate of participation decreased markedly over time and ceased completely after 6 weeks. As it turned out, 147 was just enough for four stimuli as well as the control condition. An additional stimulus would have needed around 25 new subjects; that proved impossible under the conditions of our study.[3]

Despite this constraint, we selected the four stimuli to be as diverse as possible:

- Baby crying was a continuous, irregular, human-generated sound

[3]Of course, recruitment posters could have distributed better around campus. However, experience shows that attracting participants from farther afield is harder than from nearby locations, especially given the relatively meager participation reward.

- Helicopter rotors was a continuous, regular, mechanical sound
- Hammering was a discrete, regular mechanical sound, and
- Circular saw was a continuous, irregular mechanical sound

The most obvious discrete, human-generated stimulus – talking – was intentionally omitted, since it would have likely caused confusion between the experiment instructions and the stimulus.

### 2.6.3 Insufficiently Security-Critical Task

We suspect that most participants were unaware, ahead of time, of the purpose and details of our experiment. However, during the experiment they clearly understood that the task at hand was Bluetooth-based pairing of their smartphone with some other (our) device. Consequently, from the participant's perspective, this task was unlikely to be perceived as being truly security-critical; the device the subjects were asked to connect to was obviously a prop, not a device the subject owned.

In the same vein, device pairing is neither as security-critical nor as pervasive (or frequent) as other tasks, such as password or PIN entry for the purpose of Internet access or PIN entry into an Automated Teller Machine (ATM). However, experimenting with these more natural tasks is significantly more difficult.

### 2.6.4 Synthetic Environment

Our unattended experiment setup is clearly very synthetic, for several reasons: First, it is normally very quiet, in contract with many (perhaps most) common everyday settings. Second, it is located indoors with no exposure to daylight, no air movement and no temperature fluctuations. Third, the setup (as shown in Figure 3.1) involves equipment that an average participant never or rarely encounters in the real world, in particular, a touch-sensitive Smartboard used as a means of both input and output, and a unusual-looking companion projector.

### 2.6.5 Ideal Setting

Based on the above discussion, it is easy to see that the ideal setting for our experiment would be one where:

- Demographics of participants is widely varied
- Participants are completely unaware of the experiment, at least until it is over
- The environment is common/natural
- The task is truly security-critical

One trivial example of such an ideal setting is a bank ATM located in a well-trafficked public space, with the security-critical task being the PIN entry process.

A modern ATM incorporates all features needed for our type of experiments: a keypad, a screen, a speaker (for visually impaired individuals), and a video camera. A similar setting is encountered in some automotive gas stations where the fuel pump includes a keypad (used for PIN and/or Zip code entry), a screen and a speaker; video cameras are usually located overhead. Yet another example would be a setting with public Internet access terminals, commonly found in airports and hotels, where the security-critical task would be the log-in process to the Internet provider.

In theory, in any of the above examples, large numbers of diverse subjects can be seamlessly gathered without any explicit recruitment, awareness of the experiment or reward for participation. However, it is easy to see that conducting experiments in these ideal settings would be physically, logistically and ethically problematic.

## 2.7 Ethical Considerations

Experiments described in this chapter were fully authorized by the Institutional Review Board (IRB) of our university, well ahead of the actual commencement of the study. The level of review was: Exempt, Category II. Further IRB-related

details are available upon request. We note that no sensitive data was harvested during the experiments and minimal identifying information was retained. In particular:

- As part of Bluetooth device pairing, participants were not asked to select any secret PINs or passwords. Instead, the 6-digit PIN was generated on the computer hidden from view and displayed on the Smartboard as well as their smartphone; they were then asked to compare the two PINs and confirm that they were identical.
- The hidden computer (iMac) used for pairing was periodically flushed of all collected device pairings.
- No names, addresses, phone numbers or other identifying information was collected from the participants.
- Although email addresses were solicited in order to deliver the participation reward, they were erased very soon thereafter.
- Video recordings of the experiments were (and still are) kept for study integrity purposes. However, we plan is to erase them before IRB expiration time.

Finally, with regard to safety, we maintained noise levels of between $67$ and $80$ dB which is (especially for a very short duration, i.e., less than a minute) generally

considered safe for people, as discussed earlier in Section 5.2.2.

## 2.8 Conclusion

As the "human link" in security-critical tasks becomes more popular in various settings, including those subject to accidental or adversarial sensory input, a thorough evaluation of usability in the context of such tasks becomes imperative. This work took the first step by studying the effects of unexpected audio noise on users performing wireless device pairing.

Our use of an unattended experiment led to several complications that we had not anticipated. For example, we were often confronted with multiple subjects simultaneously taking part in the experiment or advising one another on how to correctly complete the task at hand. A technical solution to this problem would be an enclosed experimental area whose access is restricted to entry by a single person only (e.g., through a controlled turnstile). Unfortunately, this would be in violation of fire safety regulations. We therefore plan to explicitly instruct participants that the experiment is intended to be conducted by a single subject at a time, and to verify and penalize non-compliance, e.g., by denying reward to non-compliant subjects.

Alternatively, instead of discarding these results, in future studies it may be worthwhile to consider and compare the results of those collaborating subjects' trials in the context of all trials with multiple participants. Such a comparison was beyond the scope of our initial experiment.

To proactively discourage multiple experiments by the same subject with different Bluetooth devices we could explicitly advertise the fact that video recordings will be reviewed and subjects who participate more than once will not receive a reward. However, this would accentuate the fact that subjects are on camera, which could potentially influence performance.

We feel that this experimental paradigm is valuable and deserves further evaluation. One possible goal is to create a new standard whereby large experiments with hundreds of subjects can be conducted without posing a prohibitive financial and/or logistical burden.

# Chapter 3

# Effects of Visual Distractions on Completion of Security Tasks

## 3.1 Introduction

It is widely believed that the human user is the weakest link in the security chain. Nonetheless, human participation is unavoidable in many security protocols. Such protocols require extensive usability testing, since users are unlikely to perform well when faced with overly difficult or intricate tasks. Typically, security-related usability testing entails evaluating human performance in a "best-case" scenario. In other words, testing is usually conducted in sterile lab-like environments.

At the same time, security protocols involving human users have become more commonplace. Examples include activities, such as: (1) using a personal device for verification of transaction amounts, (2) entering a PIN or a password and (3) solving a CAPTCHA, (4) comparing PINs when pairing Bluetooth devices, and (5) answering personal security questions. Since overall security of these tasks is determined by the human user (as the weakest link), extensive usability studies have been conducted. They aimed to assess users' ability to perform security tasks correctly and without undue delays, while providing an acceptable level of security [35] [23] [47] [29]

However, the focus on maximizing successful protocol completion led developers to evaluate usability under contrived and unrealistic settings. In practice, security tasks can take place in noisy environments. In real-world settings, users are often exposed to various sensory stimuli. The impact of such stimuli on performance and completion of security tasks has not been well studied. A particular stimulus (e.g., a fire alarm or flickering lights) can be unintentional or hostile, i.e., introduced by the adversary that controls the physical environment. Furthermore, recent emergence of Internet of Things (IoT) devices (such as smart speakers and light fixtures) in home and office settings creates environments where compromised (malware-infected) devices can expose users to a variety of visual and audio

stimuli.

There has been just one prior study that studied the effects of audio stimuli on the completion of security-critical tasks, described in the previous chapter. This initial result motivates a more thorough study in order to fully understand the effects of a range of unexpected (and potentially malicious) stimuli on user performance of security-critical tasks.

Since modern user-aided security protocols focus on maximizing successful outcomes in an ideal environment, human errors are quite rare. For example, Uzun et al. [60] assume that :

> "...[A]ny non-zero fatal error rate in the sample size of 40 is unacceptable for security applications."

Consequently, numerous trials with many subjects are needed to gather data sufficient for making claims about human error rates. The scale is further exacerbated by the need to test multiple modalities, each with a distinct set of subjects. (This is because a given subject is less likely to make a similar mistake twice, even under different conditions.) Therefore, the number of required participants can quickly grow into hundreds, which presents a logistical challenge. To ease the burden of conducting a large-scale study, we designed and employed an entirely unattended and automated experimental setup, wherein subjects receive recorded instructions

from a life-sized projection of a video-recorded experimenter ("avatar"), instead of a live experimenter. We extensively experimented with subjects attempting to pair two Bluetooth devices (one of which was the subject's own device) in the presence of various unexpected visual stimuli. We tested a total of 169 subjects in the fully unattended experiment setting.[1] We initially hypothesized that visual stimuli would have beneficial or facilitatory effects on subject task completion, as was recently experienced with its audio counterpart [28]. Surprisingly, we discovered a marked slowdown in task completion times across the board, and lower task success rates under certain stimuli.

The rest of the chapter is organized as follows: The next section presents the design and setup of our experiments, followed by the presentation of our experimental results. After that, we derive conclusions and summarize lessons learned. The chapter concludes with the discussion of limitations of our approach.

## 3.2 Methodology

This section describes our experimental setup, procedures and subject parameters.

[1] All experiments described in this chapter were fully authorized by the Institutional Review Board (IRB).

Figure 3.1: Experimental environment: (a) front view and (b) side view

### 3.2.1 Apparatus

The experimental setting was designed to facilitate fully automated experiments with a wide range of sensory inputs. Because of this, we decided to locate it in a public, but low-traffic alcove at the top floor of the main Computer Science Department building in a large public university. Figure 1(a) shows our setup from the subject's perspective (front view), and Figure 1(b) depicts it from the side. The setup is comprised entirely of the following readily available off-the-shelf components:

- A 60"-by-45" touch-sensitive interactive Smartboard[2] whiteboard with a Hitachi CP-A300N short-throw projector[2]. The Smartboard acts as both

an input and a display device. It reacts to tactile input, i.e., the user touches its surface, similar to a large touch-screen.

- A Logitech C920 HD Webcam[2].
- Two pairs of BIC America RtR V44-2 speakers[2]: one alongside the smartboard, and the other – on the opposite wall. Their arrangement is such that the subject is typically standing in the center of the four speakers.
- Four programmable wirelessly controllable Phillips Hue A19 LED lightbulbs[2] to deliver the visual stimuli.

The final component of the experiment was the subject's own Bluetooth-capable device. All prospective subjects were explicitly informed, during recruitment, that they would need to use their own personal device that supports Bluetooth communication. We could have instead provided our device, which might have fostered a more uniform subject experience. However, there would have been some drawbacks:

- We wanted to avoid accidental errors due to the use of an unfamiliar device that might have a different user interface from that of the subject's own device. Mitigating this unfamiliarity would have required some training, which is incompatible with the unattended experiment setting.

[2]See: `meethue.com` for Hue Bulbs, `smarttech.com` for the Smartboard, `logitech.com` for the Webcam, `bicamerica.com` for speakers, and `hitachi.com` for the projector.

- Virtually all current Bluetooth pairing scenarios involve at least one of the devices being owned by the person performing the pairing. Forcing subjects to use our device would have resulted in a more contrived or synthetic experience.
- From a purely practical perspective, an unattended portable device provided by us would have been more prone to damage or theft than other components, which are bulky and attached to walls and/or ceilings.

Not surprisingly, the majority of subjects' devices (152 out of 169) were smartphones. Tablets (13) and laptops (4) accounted for the rest.

Bluetooth pairing is not as common as other security-critical tasks, such as password entry or CAPTCHA solving. However, we believe that Bluetooth pairing is the ideal security-critical task for the unattended experiment setup. It is preferred to passwords and PINs since it does not require subjects to reveal existing, or to select new, secrets. The security task at the core of Bluetooth pairing involves the user comparing two 6-digit decimal numbers – one displayed by each device being paired – and pressing a single button. This is a much more discrete and uniform activity than solving CAPTCHA-s, which vary widely in terms of difficulty and require higher-resolution displays as well as more extensive user input. These factors, even without external stimuli, would yield large variations

in error rates and completion times.

### 3.2.2 A Few Colorful Words

This subsection fully defines the color system and lighting-specific terms used in the creation and evaluation of the visual stimuli.

#### Munsell Color System

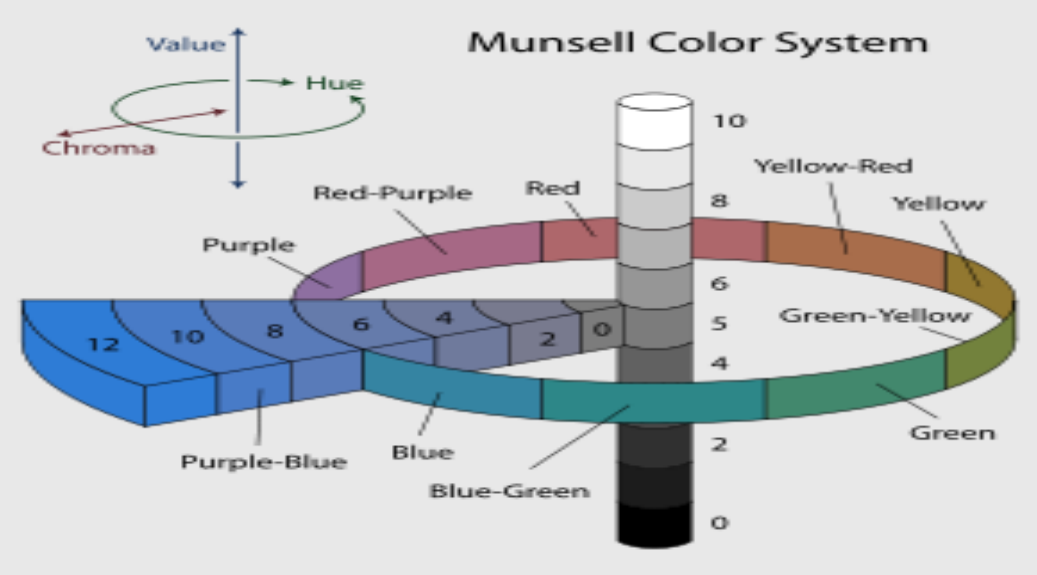


Figure 3.2: Munsell Color Space (Image best viewed in color)

The Munsell Color System is used for creating and describing colors. In it, all colors are grouped into two categories: primary and intermediate hues. Primary hues include: Red, Yellow, Purple, Blue, and Green, arranged in a circular shape

as in Figure 3.2. Intermediate hues are mixtures of two adjacent primary hues, such as Yellow-Green or Purple-Blue. Colors are defined on three dimensions: hue, lightness, and color purity. The Munsell system is based on human perception which makes it useful for rigorously defining human reaction to specific color forms. However basing the system on human perception makes the Munsell system a poor tool for direct conversion of light described by its physical wavelength into human-perceptible color.

### CIE Color Space

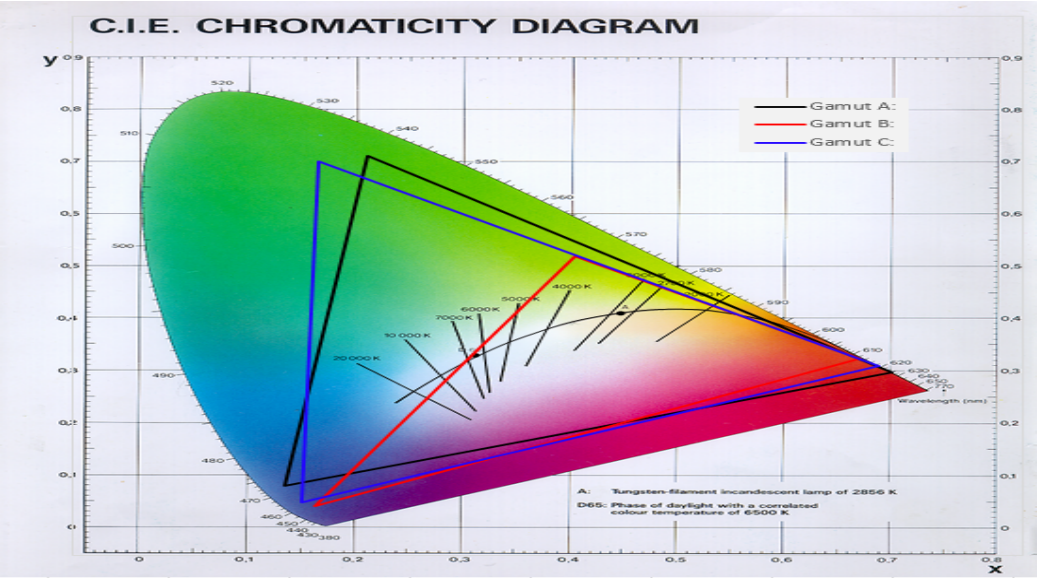


Figure 3.3: Phillips Hue CIE Color Space (Image best viewed in color)

The Phillips Hue bulbs use the CIE color space, instead of the Munsell Color

System. In CIE, colors are defined as a 2-dimensional space with X and Y values moving along a roughly triangular curve that corresponds to the translation of wavelengths of light to their human perception in the visible spectrum. The exact color range of the Philips Hue bulb is shown in Figure 3.3

**Lumens**

The lumen (lm) is the SI unit for luminous flux. It measures the total amount of visible light emitted by a source. It is defined as $1$ lm $= 1$ candela $\cdot$ steradian, where one candela approximates the luminous intensity of a single candle, and $4\pi$ steradian corresponds to a full sphere.

### 3.2.3 Procedures

As mentioned earlier, instead of a live experimenter, we used a life-size video/audio recording of a experimenter giving instructions. This avatar is the subjects' only source of information about the experiment. Actual experimenter involvement is limited to strictly off-line activities, such as infrequent recalibration of avatar video volume and visual effects, as well as occasional repair of some components that suffered minor wear-and-tear damage throughout the study. This unattended setup allows the experiment to run without interruption 24/7 over a 5

month period.

Recall that the central goal of the experiment is to measure performance of subjects who attempt to pair their personal Bluetooth device to our Bluetooth device – an iMAC that uses the SmartBoard as an external display. This iMAC is hidden from the subject's view; it is situated directly on the other side of the SmartBoard wall in a separate office. During the pairing process, each subject is exposed to one randomly selected (from a fixed set) visual stimulus. This is done by rapid change in the ambient lighting of the room's four overhead lightbulbs to the chosen stimulus condition.

The experiment runs in four phases:

1. Initial: the subject walks in, presses a button on the wall which activates the experiment. Duration: instant.
2. Instruction: the avatar delivers instructions via Smartboard display and speakers. Duration: 45 seconds.
3. Pairing: the subject attempts to pair personal device with SmartBoard which represents the hidden iMAC desktop. In this phase, the subject is exposed to one (randomly selected out of 7) visual distraction stimulus. Duration: up to 3 minutes.
4. Final: the subject is prompted, on the SmartBoard, to enter some basic

demographic information, as well as an email address to deliver the reward – an Amazon discount coupon. The information is entered directly into the SmartBoard, acting as a touch-screen input device. Duration: up to 6 minutes.

The total duration of the experiment ranged between 5 and 10 minutes.

In order to mitigate any disparities in task completion times between subjects that already had Bluetooth Discovery enabled and those who did not, the avatar informs subjects in the first 15 seconds of the instruction dialog that they will need to perform Bluetooth pairing with their personal device. This gives subjects over 30 seconds to enable Bluetooth Discovery Mode on thier device, if it is not enabled already.

We selected 6 visual effects that differed across two dimensions: color and intensity. In terms of color, we picked 3 values in the CIE chromatic space: Red, Blue, and Yellow-Green. Each is either *Solid*, i.e., shown at constant maximum intensity for the duration of the effect, or *Flickering*, i.e., its intensity grows and shrinks from the minimum to the maximum and back, completing one full cycle every second. In all settings, the maximum saturation was used. Color and intensity parameters for the 4 Phillips Hue bulbs under each condition are as follows (CCV stands for CIE Chromatic Value) [66]:

1. Red, CCV: X= 0.674, Y = 0.322
2. Blue, CCV: X = 0.168, Y = 0.041
3. Yellow-Green, CCV: X = 0.408, Y = 0.517
4. Solid intensity lumen output: 600 lm
5. Flickering intensity lumen range: 6 lm - 600 lm

These color conditions were picked based on capabilities of programmable bulbs as well as background knowledge about emotive effects of color. Phillips Hue is an LED system based on creating white light. It can not create a blacklight effect or any achromatic light, which limits color selection to the subspace of the CIE color space [66] that Hue supports.

Figure 3.4: The Subject's Perspective, Under the Red Condition.

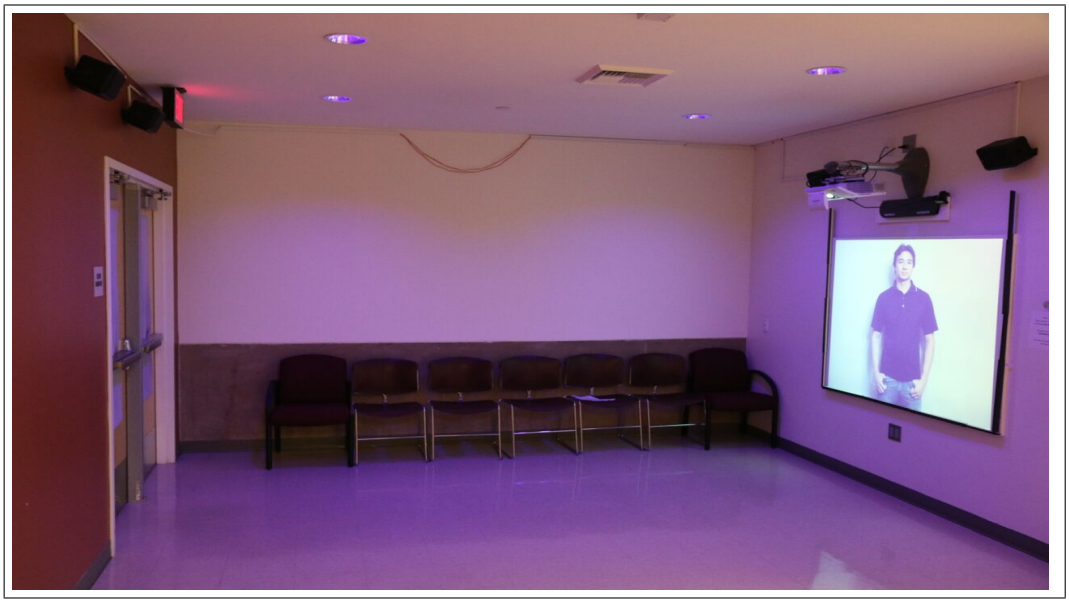

Figure 3.5: The Experimental Enviroment, Under the Blue Condition

With that restriction, we looked to the state-of-the-art about emotive reception and sensory effects of various colors in the Munsell color space [46]. It has been shown that *principal hues* – Red, Yellow, Purple, Blue, and Green – are typically positively received. In contrast, *intermediate hues*, i.e., mixtures of any two principal hues, are more often negatively associated. Also, various colors have been shown to have either an arousing or a relaxing effect on subjects exposed to them. Based on this information, we chose three colors that differ as much as possible [44]:

- Red: Principal hue with positive emotional connotations, high associated arousal levels, see Figure 3.4

- Blue: Principal hue with positive emotional connotations, low associated arousal levels, see Figure 3.5.
- Yellow-Green: Intermediate hue with negative emotional connotation, high associated arousal levels

Furthermore, we chose to have multiple modalities of light intensity for each color, with the expectation that a more complex modality would be more arousing and have a greater effect than its simple counterpart [36]. Not having found any previous work on the impact of exposure to colored light on performance of security-critical tasks, we include *Solid* light – the simplest modality of exposure that corresponds to the base level of stimulation. As a more complex modality, we included *Flickering* light.

Clearly, these two modalities were not the only possible choices. For example, it might have been intuitive to include even a more complex and startling *Strobing* light modality, achievable through rapid modulation of light intensity. It would have probably engendered a more profound impact on the subjects. However, ethical considerations coupled with the unattended nature of the experiment preclude using any modality that could endanger subjects with certain sensitivity conditions, such as photosensitive epilepsy. This led us to select a safe flickering frequency of $1Hz$.

We also found that all three light colors (under both intensity modalities) do not interfere with readability of a backlit personal wireless device or the image projected on the Smartboard. All experimenters, including one who used corrective lenses, could *correctly* read the screens of their personal devices, under all color conditions and intensity modalities.

### 3.2.4 Initial Hypotheses

We started out by hypothesizing that introduction of unexpected visual distractions during the process of human-aided pairing of two Bluetooth devices would have similar effects to those observed in prior experiments with audio distractions. Specifically, we expected two outcomes, as compared to a distraction-free setting:

**[H1]:** Lower error rates, and

**[H2]:** No effect on task completion times

### 3.2.5 Recruitment

The main challenge we encountered in the recruitment process is the scale of the experiments. Prior studies of usability of human-aided pairing protocols [23, 47, 29], demonstrated that 20-25 subjects per tested condition represents acceptable size for obtaining statistically significant findings. Our experiment has one con-

dition for each of the six visual distraction variations, plus the control condition with no distractions. Therefore, collecting a meaningful amount of data requires at least 140 iterations of the experiment.

We used a four-pronged strategy to recruit subjects:

1. Email announcements sent to both graduate and undergraduate Computer Science students.
2. Posters placed (as signboards) near the entrance, and in the lobby, of a large campus building which housed the experimental setup.
3. Several instructors promoted participation in the experiment in their lectures.
4. Printed fliers handed out at various campus locations during daily peak pedestrian traffic times.

Recruitment efforts yielded 169 subjects in total, of whom 125 were male and 44 – female, corresponding to a 74%-26% gender split. This is expected, given that the location of our experimental setup was in the Computer Science and Engineering part of campus. Most subjects (161) were of college age (18-24 years), while 8 were in the 30+ group. This distribution is not surprising given the university population and the fact that older subjects generally correspond to researchers, faculty and staff, all of whom are much less likely to be attracted to

being a subject in an experiment.

As follows from the above, our subjects' demographic was dominated by young, tech-savvy male undergraduate students.

## 3.3 Results

This section discusses the results, starting with data cleaning and proceeding to subject task completion effects.

### 3.3.1 Data Cleaning

We had to discard subject data for three reasons.

First, although instructions (in fliers, announcements, and signs near the setup) specifically stated that subjects were to arrive alone, and perform the experiment without anyone else present, 37 groups (2 or more) of subjects participated. We found that the initial participant from each group performed in a manner consistent with individual subjects. However, subsequent group members who tried the experiment were (not surprisingly) significantly faster and more accurate in their task completion. Consequently, we discarded data of every subject who arrived in a group and was not the initial participant.

Table 3.1: Subject Failure Statistics

| Stimulus | #Successful Subjects | #Failed Subjects | Failure Rate |
|---|---|---|---|
| None (control) | 32 | 15 | 0.32 |
| Solid Red | 11 | 9 | 0.45 |
| Flickering Red | 9 | 11 | 0.55 |
| Solid Blue | 14 | 6 | 0.30 |
| Flickering Blue | 8 | 12 | 0.60 |
| Solid Yellow-Green | 10 | 12 | 0.54 |
| Flickering Yellow-Green | 7 | 13 | 0.65 |
| **Total** | 91 | 78 | 0.46 |

Table 3.2: Barnard's Exact Test on failure rates

| **Stimulus** | Total Pairings | Failure Rate | Wald Statistic | Nuisance Parameter | $p$ |
|---|---|---|---|---|---|
| None(Control) | 47 | 0.32 | – | – | – |
| Solid Red | 20 | 0.45 | 1.02 | 0.88 | 0.17 |
| Flickering Red | 20 | 0.55 | 1.77 | 0.86 | 0.04 |
| Solid Blue | 20 | 0.30 | 0.15 | 0.05 | 0.49 |
| Flickering Blue | 20 | 0.60 | 2.14 | 0.96 | 0.03 |
| Solid Yellow-Green | 22 | 0.54 | 1.79 | 0.94 | 0.06 |
| Flickering Yellow-Green | 20 | 0.65 | 2.51 | 0.91 | 0.01 |

The second reason for discarding data would have been due to subject auditory and/or visual impairment. A subject with an auditory impairment would have difficulties understanding the avatar's spoken instructions. A visually impaired subject would have difficulties with using the Smartboard and with the pairing process which relies on reading and comparing numbers. After carefully reviewing all subject video records, we could not identify any obvious visual or auditory impairment in any subject.

Some subjects successfully completed the experiment several times, perhaps

hoping to receive multiple participation rewards. This occurred despite explicit instructions to the contrary. The system automatically rejected any repeated pairing attempts from devices already paired with the system, and any repeated attempts with different devices were discovered by visual inspection of subject trials. Every such repeated instance was discarded.

Finally, each subject was exposed only to a single color condition and was not required to distinguish between multiple colors. Because of this, color-blindness should have minimal impact on subject results and we did not vet for it. This decision is supported after-the-fact by results of the one-way Analysis of Variance test described later on in this section.

### 3.3.2 Task Failure Rate

Table 5.1 shows the number of subjects who, respectively, succeeded and failed at Bluetooth device pairing under each stimulus condition. It also details the failure rate for each condition.

Table 3.2 shows results from Barnard's exact test applied pairwise to the subject failure rate of the control condition and each stimulus. It demonstrates that differences between failure rates are statistically significant at the $\alpha = 0.05$ level with respect to all *Flickering* conditions: *Flickering Red*, *Flickering Blue*, and

*Flickering Yellow-Green*. This even holds if we apply a conservative Bonferroni correction to account for three pairwise comparisons. This leads us to the mixed rejection of the initial hypothesis **H1**, as the failure rate increases significantly with the introduction of certain kinds of visual distractions, and remains unaffected by others. The next section discusses this further.

Table 3.3 shows odds ratios and 95% confidence interval for the failure rates under each stimulus, as compared to the control condition's failure rate. Interestingly, under this analysis, only the confidence intervals of *Flickering Blue* and *Flickering Yellow-Green* do not include a possible odds ratio of $1.0$. Therefore – under this method of analysis – they are the only statistically significant stimuli at the $\alpha = 0.05$ level. The confidence interval defined for the *Flickering Red* condition challenges the claim of statistical significance at the $\alpha = 0.05$ level, as established by Barnard's exact test.

We also examined subject failure rates by gender. As shown by Table 3.4 there is no statistically significant difference in failure rates between male and female participants; Wald statistic $= 0.36$, nuisance parameter = $0.01$, $p = 0.46$.

Table 3.3: Subject Failure Rate by Gender

| Stimulus | Odds Ratio wrt control | 95% Confidence Interval wrt control |
|---|---|---|
| None (control) | - | – |
| Solid Red | 1.70 | 0.60-5.11 |
| Flickering Red | 2.61 | 0.89-7.63 |
| Solid Blue | 0.91 | 0.29-2.85 |
| Flickering Blue | 3.20 | 1.08-9.47 |
| Solid Yellow-Green | 1.79 | 0.91-7.24 |
| Flickering Yellow-Green | 3.96 | 1.31-11.6 |

Table 3.4: Subject Failure Rate by Gender

| Gender | #Successful Subjects | #Unsuccessful Subjects | Failure Rate |
|---|---|---|---|
| Male | 65 | 59 | 0.48 |
| Female | 25 | 20 | 0.44 |

### 3.3.3 Task Completion Times

Table 5.2[3] [4] shows average completion times in successful trials under each stimulus. After applying a conservative Bonferroni correction to account for six pairwise comparisons between individual stimulus conditions and the control condi-

[3]Std Dev = Standard Deviation

[4]DF = Degrees of Freedom.

tion, every stimulus condition shows an overwhelmingly large, statistically significant departure from the control condition. This results in rejection of hypothesis **H2**. The following section examines possible causes of this slowdown, as well as its implications.

Table 3.5: Avg times (sec) for successful pairing.

| Stimulus | Mean Time | Std Dev | DF wrt control | t-value wrt control | $p$ |
|---|---|---|---|---|---|
| None | 34.50 | 11.93 | – | – | – |
| Solid Red | 87.81 | 24.56 | 41 | 9.56 | $< 0.001$ |
| Flickering Red | 90.44 | 15.62 | 39 | 11.59 | $< 0.001$ |
| Solid Blue | 106.36 | 17.39 | 44 | 16.32 | $< 0.001$ |
| Flickering Blue | 91.25 | 24.11 | 38 | 9.61 | $< 0.001$ |
| Solid Yellow-Green | 90.30 | 19.08 | 40 | 11.1 | $< 0.001$ |
| Flickering Yellow-Green | 90.29 | 19.06 | 37 | 10.01 | $< 0.001$ |

Table 3.6 shows Cohen's $d$ for completion times under each stimulus when compared to the control condition. $|d| > 1.0$ in all cases, which means that every stimulus condition shows an overwhelmingly large, statistically significant departure from the control condition for the evaluation of Cohen's $d$. This result is

Table 3.6: Cohen's $d$ on Completion Times wrt Control

| Stimulus | Cohen's d wrt control |
|---|---|
| None (control) | - |
| Solid Red | -3.42 |
| Flickering Red | -4.49 |
| Solid Blue | -5.33 |
| Flickering Blue | -3.90 |
| Solid Yellow-Green | -4.12 |
| Flickering Yellow-Green | -4.29 |

statistically significant: it indicates that, with convincing probability, the mean completion time observed under the control is representative of a different distribution than that observed under each stimulus condition. This supports rejection of hypothesis **H2**.

Next, we looked into subject completion times for successful completion attempts by gender. Results are displayed in Table 3.7. A pairwise t-test shows that observed differences are not statistically significant; $t(84) = 0.04$, $p = 0.96$.

Finally, we preformed Bartlett's test for homogeneity of variances as well as a One-Way analysis of variance (ANOVA) test between average task completion times of all stimulus conditions, excluding the control. Bartlett's test failed to

Table 3.7: Avg times (sec) by gender

| Gender | Mean Time | Standard Deviation |
|---|---|---|
| Male | 75.27 | 22.31 |
| Female | 75.20 | 24.10 |

Table 3.8: One-Way ANOVA test

| | Sum of Squares | DF | Mean Square | $F$ | $p$ |
|---|---|---|---|---|---|
| Between Groups | 2964.28 | 5 | 592.86 | 1.466 | 0.217 |
| Within Groups | 21440.33 | 53 | 404.535 | | |
| Total | 24404.61 | 58 | | | |

reject the null hypothesis that all stimulus conditions share the same variance ($\chi^2 = 2.80$, $p = 0.731$). Furthermore, the one-way ANOVA test indicated no significant difference between any sample distributions (F = 1.466, p = 0.217.) Table 3.8shows the results; their implications are discussed in the following section.

Table 3.9: Failure Rates: group initiators vs. individuals

| Participant Type | #Successful Subjects | #Unsuccessful Subjects | Failure Rate |
|---|---|---|---|
| Group Initiator | 19 | 18 | 0.49 |
| Individual | 72 | 60 | 0.45 |

Table 3.10: Avg times (sec): Group Initiators vs. Individuals

| Participant Type | Mean Time | Standard Deviation |
|---|---|---|
| Group Initiator | 76.63 | 23.00 |
| Individual | 76.20 | 17.93 |

### 3.3.4 Analysis of Group Initiators

We considered potential differences in failure rates between subjects who performed the task alone (per instructions), and those who did it as part of a group. As mentioned in the discussion of Data Cleaning, for each group, we only consider the initial participating group member, referred to as the Group Initiator. As Table 3.9 shows, there is no significant difference between failure rates of individual subjects and Group Initiators; Wald Statistic $= 0.34$, Nuisance parameter $= 0.01$, $p = 0.51$. Furthermore, as Table 3.10 shows, a pairwise t-test of completion times for individuals – compared to group initiators – shows that observed differences are not statistically significant; $t(84) = 0.09$, $p = 0.93$.

## 3.4 Discussion of Observed Effects

Several types of visual stimuli appear to have a negative effect on the subjects' successful completion of the Bluetooth Pairing task. However, collected data shows that this is not consistent across all stimuli. Instead, the negative effect may be tied to certain features of the particular stimulus. Instances of significant degradation in subject success rates were linked to the *Flickering* modality, for all color stimuli. This result implies that emotional perception of the stimulus may not be as much of a contributing factor to the overall increase of subject arousal as the presence of a dynamic visual stimulus. Also, in contrast with a previous study of audio distractions that observed positive effects [28], we noted no benefit to subject success rates under **any** visual stimulus.

These negative and neutral responses to static and dynamic light stimuli, respectively, are reinforced by the psychological concept of attentional selectivity. This concept assumes that the capture of an individual's attention by an aversive stimulus is likely to be momentary, occurring primarily when the stimulus is first introduced. In cognitive science, attention is conceptualized as a limited resource. For good evolutionary reason, the greatest demand on attention is in response to any change in one's environment. Once an assessment of the stimulus is made, and determined not to require additional action, attentional devotion to that stim-

ulus fades quickly. This means that – while a static, adverse lighting change may remain adverse throughout its duration – its capacity to interfere with subject performance will fade rapidly after its onset. Instead, dynamically changing stimuli can more effectively capture subject attention and impair their performance, since many assessments are needed for many environmental changes occurring throughout the stimulus's duration.

Negative impact on subject task completion rates prompts a new attack vector for the adversary who controls ambient lighting. By taking advantage of color effects with shifting intensity levels, the adversary could force a user into failing Bluetooth pairing as a denial-of-service (DoS) attack. Moreover, the adversary might induce failure by using positively perceived colors of varying intensity. These colors may not even register as malicious in the user's mind, as they are innately associated with beneficial or pleasant emotions.

However, a much greater effect was observed in terms of average completion time. During review of subject trials, we noted that, upon exposure to the stimulus, subjects often take their gaze off their personal device (or the avatar) and focus their attention to the colorful, and possibly flickering, lights. The resulting delay frequently caused the subject's device to exit the Bluetooth pairing menu due to a time-out, and re-initiate the pairing protocol, resulting in much longer completion

times overall.

Furthermore, as shown by Table 3.8, the introduced delay in subject task completion time was not based on the particular stimulus. Instead, the mere presence of a visual stimulus was enough to slow down successful subjects. Similar to the result in inducing user failure, the adversary is not forced to rely on an overtly malicious stimulus in order to cause substantial slowdown in task completion. However, the adversary has even more choices in stimulus selection, since all stimuli (including those with static intensity levels) were shown to impact task completion times the same way.

This effect shows further power for the adversary in control of ambient lighting. One possibility is that the adversary's goal is a denial-of-service attack by frustrating user's pairing attempts. In a more sinister scenario the adversary could try to "buy time" by introducing its own malicious device(s) alongside changes to ambient lighting and then leverage the user's lapse in focus (when being exposed to new sensory stimuli) to trick the user into pairing with that device. In the worst case, the adversary might take advantage of the user's inattentiveness while their gaze shifts away from their device and trick them into accepting a non-matching authenticator.

## 3.5 Unattended Setup: Limitations

Based on our earlier discussion of Data Cleaning, some subjects' data had to be removed from the dataset because they did not conduct the experiment alone. This occurred even though all recruitment materials (and means) as well as the avatar's instructions stated that subjects were to perform the task alone. This illustrates a basic limitation of the unattended setup: no one is present to enforce the rules in real time.[5]

We did not manage to capture fine-grained data about the subjects' awareness of a distraction. We have some anecdotal evidence from video recordings showing that some subjects noticed the distraction in obvious ways, e.g., verbal remarks or turning their heads. However, we have no evidence of subjects who failed to notice the stimulus. Information about subjects noticing a change in the environment is very important to the development of a realistic adversary model for future studies.

## 3.6 Study Shortcomings

In this section we discuss some shortcomings of our study.

[5]However, it would have been possible (though quite difficult in practice) to instrument our recording of the experiment to abort upon detecting simultaneous presence of multiple subjects.

### 3.6.1 Homogeneous Subjects

Our subject group was dominated by young, tech-savvy male college students. This is a consequence of the experiment's location. Replication of our experiment in a non-academic setting would be useful. However, recruiting a really diverse group of subjects is hard. Ideal venues might be stadiums, concert halls, fairgrounds or shopping malls. Unfortunately, deployment of our unattended setup in such public locations is logistically infeasible. Since these public areas already have many sensory stimuli, reliable adjustment of our subjects' arousal level in a consistent manner would be very hard. Furthermore, it would be very difficult to secure specialized and expensive experimental equipment.

In addition to being tech-savvy, young subjects are in general more apt to quickly recover from changes in the lighting of their surroundings than older adults [33]. It is possible that unexpected visual stimuli would have a different effect on an older (less technologically adept) population.

### 3.6.2 Sufficiently Diverse Stimuli

We selected six conditions to obtain as many diverse stimuli types as we could rigorously test, in addition to control. We first varied them by changing the regularity of the stimulus, expecting that a varying signal would have greater impact

on subjects' arousal than a steady signal. We then varied the colors, with the expectation that using colors that evoked different emotive responses and general arousal levels would impact task performance differently.

An ideal experiment would have included a stimulus with negative emotional connotation and low arousal levels. However, between three colors, two intensity conditions, and the control, we had seven total conditions to test. Furthermore, due to the nature of our experiment, we could only reasonably expect each subject to be tested under a single condition, since prior knowledge about the experiment would clearly bias the results. Adding just one additional stimulus (for both intensity modalities) would have required at least 40 more subjects. This would have placed a heavy logistical burden for our already nearly-depleted subject pool.

We also note that variance in intensity of our flickering modality did not approach the technical limit of Philips Hue bulbs. Instead, we deliberately limited the frequency of intensity fluctuations to $1Hz$ in order to avoid any possible negative reaction from light-sensitive subjects. This ethical issue does not reflect real-world conditions where an adversary (with no ethical qualms) could create a very fast strobing effect, possibly causing physical harm.

### 3.6.3 Synthetic Environment

Our unattended setup, while a step closer to an everyday setting than a sterile and highly controlled lab, is still quite synthetic. First, our choice to place it in a low-traffic area makes it quieter than many common settings. Second, our choice to situate it indoors makes it free of temperature fluctuations, air flow, and exposure to sunlight. Finally, our equipment (such as the Smartboard projector system) is not commonly encountered by most subjects.

### 3.6.4 Ideal Setting

Drawing upon aforementioned shortcomings, the ideal setting for our experiment would be one where:

- Subject demographics are more varied
- Subjects are not aware of the nature of the experiment until they are debriefed after task completion
- The environment is more commonplace
- The task is more security-critical

All of these criteria could be trivially met if, for example, we conducted the experiment at a busy bank ATM. The task at hand would be the obviously security-critical entry of the subject's PIN. A modern ATM comes standard with all of

the features needed for our experiment: it has a keypad, a screen, a speaker (for visually impaired users), a video camera, and are in areas that are artificially lit. Similarly, a busy gas station would fit our needs, as each fuel pump typically includes a keypad for PIN entry, speakers, a screen, artificial lighting, and a video camera recording the transaction. However, despite their attractive qualities, there would be serious ethical and logistical obstacles to setting up an unattended automated experiment in one these location examples.

## 3.7 Conclusions

As human participation in security-critical tasks becomes more commonplace, so does the incidence of users performing these tasks while subject to accidental or malicious distractors. This strongly motivates exploring user error rates and their reactions to various external stimuli. Our efforts described in this chapter shed some light on understanding human errors in security-critical tasks by studying the effects of visual stimuli on users attempting to pair two Bluetooth devices.

We feel that this unattended experiment paradigm is a valuable approach that deserves further study. The development of standardized unattended and automated experimental setups could greatly lower the logistical and financial burdens

associated with conducting large-scale user studies.

Given the observed negative effect on subject completion times, one interesting next step would be to conduct a similar experiment, where, instead of measuring subjects' ability to pair Bluetooth devices, we would examine the rates of incorrect pairing when the subjects are shown mis-matched numbers during the pairing process. This could help us determine whether (and how) visual distractions make users more likely to pair their device to some other (perhaps adversary-controlled) device.

# Chapter 4

# Exploring Effects of Auditory Stimuli on CAPTCHA Performance

## 4.1 Introduction

**C**ompletely **A**utomated **P**ublic **T**uring tests to tell **C**omputers and **H**umans **A**part (CAPTCHAs) are programs that generate and evaluate challenges that are easy for a human to solve, yet cannot be easily solved by software. CAPTCHAs have been used to prevent bot-based abuse of services for over a decade [62]. For better or for worse, they have become a fairly routine hurdle for users seeking to access online resources, such as: discussion forums, ticket sales, banking, and email

account creation.

Because of their widespread adoption, successful attacks, and dislike by users, most recent efforts in development have been invested into creating CAPTCHAs that are [13]:

1. Usable: humans are successful at least $90\%$ of the time.
2. Secure/robust: a state-of-the-art bot should not be successful more than $0.01\%$ of the time.
3. Scalable: challenges that are either automatically generated, or drawn from a body that is too large to hard-code responses for each possible challenge.

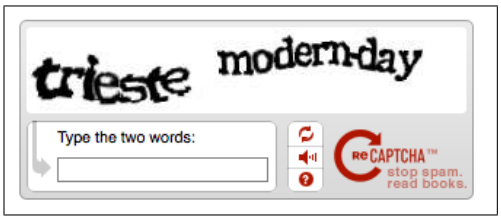


Figure 4.1: A Text-Based CAPTCHA.

Based on these requirements CAPTCHA developers focused on text-based CAPTCHAs, i.e., those that present a jumbled alphanumeric code as the challenge to the user, as shown in Figure 4.1. This approach is popular since human users are quite good at identifying these alphanumeric codes in an altered image, thus satisfying the usability requirement. Also, image segmentation and recovery is a known hard problem for AI, satisfying the security requirement. Finally, such challenges can

be randomly generated as needed, thus satisfying the scalability requirement [21].

However, scant attention has been paid to the user's physical context while solving CAPTCHAs. Security-critical tasks, such as CAPTCHAs, are often completed in noisy environments. In these real-world settings, users are often exposed to various sensory stimuli. The impact of such stimuli on performance and completion of security-critical tasks has not been thoroughly explored. Any specific stimulus (e.g., a police siren or a fire alarm) can be incidental or malicious, i.e., introduced by the adversary that controls the environment. This threat is exacerbated and accentuated by the growth in popularity of Internet of Things (IoT) devices, particularly in contexts of "smart" homes or offices. As IoT devices become more common and more diverse, their eventual compromise becomes more realistic. One prominent example is the Mirai botnet [37] which used a huge number of infected smart cameras as zombies in a massive coordinated DDoS attack.

A typical IoT-instrumented home environment, with "smart" lighting, sound and alarm systems (as well as appliances) represents a rich and attractive attack target for the adversary that aims to interfere with a user's physical environment in particular in order to inhibit successful CAPTCHA solving. We believe that this is especially relevant to some time-critical scenarios, such as web sites that sell limited numbers of coveted tickets for concerts, festivals, promotional airfares,

etc. In these settings, a delay of just a few seconds can make a very big monetary difference.

In order to explore effects of attacks emanating from the user's physical environment we experimented with numerous subjects attempting to solve text-based CAPTCHAs in the presence of unexpected audio stimuli. We tested a total of $51$ subjects in a fully unattended experimental setting. We initially hypothesized that introduction of audio stimuli would negatively impact subject task completion. While this was mostly confirmed, certain types of stimuli surprisingly demonstrated positive effects.

The rest of this chapter is organized as follows: The next section describes the design and setup of our experiments, followed by experimental results. Next, implications of the results and advantages of the unattended experimental environment are discussed. The chapter concludes with limitations of our approach and ethical considerations.

## 4.2 Methodology

This section describes our experimental setup, procedures and subject parameters.

### 4.2.1 Apparatus

Our experimental setting was designed to allow for fully automated experiments with a wide range of sensory inputs. To accommodate this, we located the experiment in a dedicated office in the Psychology Department building of a large public university. The setup is comprised entirely of the following popular commercial-off-the-shelf (COTS) components:

- Commodity Windows desktop computer with keyboard and mouse
- 19" Dell 1907FPc monitor.
- Logitech C920 HD Webcam.
- Logitech Z200 Stereo Speaker System[1].

This experimental setup is supposed to mimic the typical environment where an average user might be presented with a CAPTCHA, i.e., a private office or room.

### 4.2.2 Procedures

As mentioned earlier, the experimental environment was entirely unattended. An instructional PowerPoint presentation was used for subject instruction, instead of a live experimenter. This presentation was each subject's only source of information about the experiment. Actual experimenter involvement was limited to off-line

[1] with the volume knob physically disabled.

activities: (1) periodic recalibration of auditory stimuli, and (2) occasional repair or repositioning of some components that suffered minor damage or were moved throughout the study's lifetime. This unattended setup allowed the experiment to run without interruption 24/7/365. It was conducted over a 3-month period at the end of 2017.

The central goal was to measure performance of subjects attempting to solve as many CAPTCHAs as possible within a fixed timeframe. Subjects were expected to solve them continuously for $54$ minutes. During this period, a subject was exposed to 4 rounds of 6 auditory stimuli. The control and stimuli were presented in a random order within each round, to mitigate any ordering effects on subject performance.

**Why CAPTCHAs?** We picked CAPTCHAs as the security-critical task for several reasons. First, CAPTCHAs do not require the subjects to enter any personally identifying information (PII) or secrets in order to solve them, and can be dynamically generated on the fly, allowing for the study of subject behavior across many different solution attempts. This is in contrast with other security-critical tasks, such as password entry. Second, solving CAPTCHAs is a fairly common task and it is reasonable to assume that all potential subjects are familiar with them, as opposed to infrequent tasks, e.g., Bluetooth pairing. Finally, the cognitive effort

needed to solve CAPTCHAs (recognize-and-type) is higher than the simple comparison task in Bluetooth pairing, and is similar to recall-and-type tasks, such as password entry [54].


Figure 4.2: Sample CAPTCHA as presented to a subject.

The experiment runs in four phases:

1. **Initial:** subject enters the office, sits down at a desktop computer and starts the instructional PowerPoint presentation. Duration: Negligible.
2. **Instruction:** subject is instructed in the nature of CAPTCHAs and the experimental procedure. Duration: 2-4 minutes
3. **CAPTCHA:** subject is presented with a random CAPTCHA. Upon submitting a solution, a new CAPTCHA is presented, regardless of the accuracy of the response. Subjects are exposed to the stimulus conditions for 24 rounds, each round lasting 2:15. Duration: 54 minutes. See Figure 4.2 for a CAPTCHA example.
4. **Final: subject is taken to a survey page and asked to enter basic demo-**

**graphic information. Duration: 2-3 minutes**

The entire experiment lasts between $58$ and $61$ minutes. Each subject's participation is recorded by the webcam and by screen-capturing software, to ensure compliance with the procedure. Since our objective is to assess overall impact of auditory stimuli on subject performance (and not performance degradation due to a surprise), the first $15$ seconds of each stimulus condition were not used in data collection. This should accurately capture the enduring effect of the auditory stimuli, and ignore the spiking effect (i.e., surprise) on the attentional system due to the introduction of an unexpected stimulus [54].

### 4.2.3 CAPTCHA Generation

Since the study was concerned primarily with usability and less with robustness, we used text-based CAPTCHAs that follows the guidelines of [14] to create challenges that are highly usable, and can be quicky solved in bulk. To facilitate this, a challenge generation algorithm was selected that created 5-character alphanumeric codes with thin occluding global lines, a small amount of global distortion and minimal local distortion of the characters. This yielded challenges that our subjects could easily and quickly solve in the baseline, i.e., control case.

### 4.2.4 Stimuli Selection

The experiment consisted of two categories of auditory stimuli: (1) static with single volume level, and (2) dynamic, that changed volume throughout presentation.

Static sound stimuli were the sounds of: (1) crying baby, (2) babbling brook, and (3) human voice reading individual letters and digits in random order at a rate of two per second. (1) and (2) were chosen for their ecological significance as a source that needs attention, and a relaxing sound, respectively. The human voice stimulus was chosen to interfere with the task-specific cognitive processes used to solve CAPTCHAs. This is analogous to the Stroop effect, a phenomenon where subjects who attempt to read the written name of a color that is rendered in a different color (e.g., the word ”red” written in blue ink) do so slower and in a more error-prone way than reading the same words in plain black ink [40]. Specific volumes of the three static stimuli were:

1. Crying baby: 78 dB
2. Babbling brook: 70 dB
3. Human voice: 75 dB

The two dynamic stimuli included: (1) randomly generated looming sounds, and (2) randomly ordered menagerie of natural, aversive sounds. The looming stimulus was an amplitude modulated tone that increased from nearly silent to 85 dB

over 5 seconds. Its intensity curve is shown in Figure 5.1. Once the looming sound completed, it repeats at a different Left/Right speaker balance, selected randomly. This repeats continuously for the entire 2:15 minute stimulus window. The natural stimulus consisted of a randomly generated sequence of aversive sounds, which included: circular saw cutting wood, blaring vuvuzela, nails on a chalkboard, and spinning helicopter rotors. These sounds were played at a randomly selected volume from 75 to 88 dB. Each lasted for up to 2 seconds before changing to the next random sound.

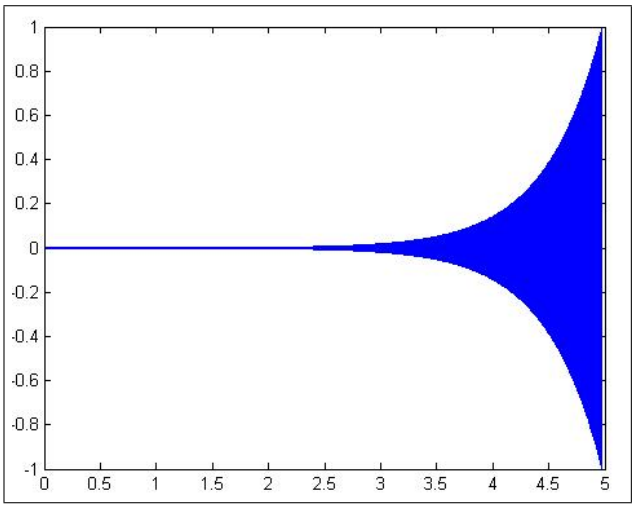


Figure 4.3: Looming Sound Intensity Function

Even the highest stimuli volume (88 dB) is well within the *safe range*, as defined by US Occupational Safety & Health Administration (OSHA) guidelines.[2].

[2]OSHA requires all employers to implement a Hearing Conservation Program where workers are exposed to a time-weighted average noise level of 90 dB or higher over an 8 hour work shift.

Clearly, an adversary that controls the victim's environment would not be subjected to any such ethical guidelines, and could thus use much louder stimuli.

### 4.2.5 Psychophysical Description of Stimuli

The chosen stimuli have the potential to produce different effects. Except for the babbling brook, selection of the sounds was guided by the intent to elicit a negative emotional response and increased level of general arousal. It is reasonable to expect a negative impact of these sounds on task performance. However, any capture of an individual's attention by an aversive stimulus is likely to be momentary, occurring primarily when the stimulus is first introduced. In cognitive science, attention is conceptualized as a limited resource. Probably for good reason, the greatest demand on attention is in response to a change in the environment. Once an assessment is made that a stimulus does not require a response, adaptation to the stimulus from a foreground target into a background context proceeds relatively rapidly as attention is redistributed to other demands. Although an aversive sound may remain aversive throughout its presentation, its capacity to disrupt performance on a complex task might rapidly fade after onset. This could serve to sharpen an individual's focus for the task at hand [56].

Our noise levels were for a much lower duration, and only the very loudest was within the regulated range. See: `https://www.osha.gov/SLTC/noisehearingconservation/`

However, the auditory attentional system is not nearly as adept at dealing with many rapid changes in the environment that occur in quick succession [7]. Dynamic synthetic sounds can be designed to attract attention resources without being aversive. To the human auditory attention system, a looming sound is not easily classified as a single, non-threatening change in the environment. Instead, it embodies a context of continuous, approaching and potentially threatening change. This unclassifiable context "tricks" the system into a state of sustained engagement, and can deplete the subject's attentional resources. Because of this phenomenon, we suspect that highly dynamic sounds have the greatest impact on subject performance.

Table 4.1: Subject Failure Rates

| Stimulus | #Successful Entries | #Unsuccessful Entries | Failure Rate | Odds Ratio wrt Control | $p$ |
|---|---|---|---|---|---|
| None (Control) | 6413 | 616 | 0.088 | - | - |
| Baby | 6074 | 1544 | 0.203 | 2.31 | $< 0.001$ |
| Brook | 6332 | 574 | 0.083 | 0.901 | 0.090 |
| Looming | 5039 | 719 | 0.125 | 1.483 | $< 0.001$ |
| Natural | 5787 | 723 | 0.111 | 1.299 | $< 0.001$ |
| Voice | 4582 | 697 | 0.132 | 1.581 | $< 0.001$ |
| **Total** | 34227 | 4873 | 0.125 | - | - |

Table 4.2: Avg Times (sec) for Successful Solutions

| Stimulus | Mean Time | Standard Deviation | DF wrt Control | t-value wrt Control | $p$ | Cohen's D |
|---|---|---|---|---|---|---|
| None (Control) | 4.621 | 3.771 | - | - | - | - |
| Baby | 4.520 | 5.267 | 12485 | 0.016 | 0.986 | 0.022 |
| Brook | 3.472 | 5.100 | 11743 | 15.026 | $< 0.001$ | 0.400 |
| Looming | 6.092 | 2.212 | 11450 | 17.373 | $< 0.001$ | 0.323 |
| Natural | 5.909 | 4.751 | 12198 | 18.505 | $< 0.001$ | 0.300 |
| Voice | 6.480 | 6.985 | 10993 | 18.07 | $< 0.001$ | 0.331 |

### 4.2.6 Initial Hypotheses

Our initial intuitive hypothesis was that introduction of unexpected auditory stimuli while solving CAPTCHAs would have negatively impact subject performance. We expected two outcomes, as compared to a distraction-free (Control) setting:

**[H1]:** Higher error rates, and

**[H2]:** Longer completion times in successful cases

We hypothesized this because, although mixed results were observed in [8] for Bluetooth pairing, solving CAPTCHAs is a more difficult cognitive task (requires more attention) even in the distraction-free (Control) case [56].

### 4.2.7 Recruitment

Recruitment was handled through the human subjects lab pool of Psychology Department at a large public university. A brief description of the study was posted on an online bulletin, and undergraduate students were allowed to sign up for the experiment and were compensated with course credit. Not surprisingly, the subject pool was dominated by college-age ($18 - 25$) individuals and the gender split was somewhat uneven: $35$ female ($69\%$) and $16$ male subjects ($31\%$).

## 4.3 Results

This section discusses the results, starting with data cleaning and proceeding to subject task completion effects.

### 4.3.1 Data Cleaning

A total of $58$ subjects took part in the study. However, $7$ of them were non-compliant with the experimental procedure, and prematurely quit the experiment. Since this behavior was captured by the recording software, all data from these subjects was discarded.

### 4.3.2 Task Failure Rate

As Table 5.1 shows, every audio stimulus – except for brook – had a substantial, statistically significant impact on subject failure rates. Furthermore, each of these was shown by their Odds ratios to have a large effect size. Thus, the impact on failure rates, though seemingly small, is a large proportional increase in failures when subjects are exposed to any stimulus, with the most impactful stimulus (crying baby) more than **doubling** subject failure rates. Interestingly, there was no direct correlation between dynamicity of the stimulus and its impact on failure rates, as the Brain Arousal Model would suggest [56]. This opens up an attack space for the adversary that controls the auditory environment, as discussed in Section 5.4.

Table 4.3: One-Way ANOVA Between Stimulus Completion Time Distributions

| Source of Variation | Sum of Squares | Degrees of Freedom | Variance | $F$ | $p$ |
|---|---|---|---|---|---|
| Between Groups | 41601.394 | 4 | 10400.349 | 412.340 | $< 0.0001$ |
| Within Groups | 676183.752 | 26809 | 25.222 | | |
| Total | 717785.146 | 26813 | | | |

### 4.3.3 Task Completion Times

Table 5.2 shows average completion times for successful CAPTCHA completions under each stimulus. Results illustrate that all stimuli (except crying baby) have a statistically significant departure from the mean ($p < 0.001$) after applying a conservative Bonferroni correction to account for $5$ pairwise comparisons to control. However, while the looming, natural and voice stimuli have a negative effect on subject performance and slow down subject task completion, brook has a positive effect and lower average task completion times. Also, although these effects appear to be highly pronounced due to their significance, their effect size is small, with Cohen's D values ranging from $0.300$ to $0.400$. Implications of these impacts on task completion times are discussed in Section 5.4.

Table 4.3 shows a one-way analysis of variance (ANOVA) evaluation of differences in means of each stimulus, excluding Control. There is a significant difference ($p < 0.0001$) in completion times across different stimuli. Furthermore, Bartlett' test for homogeneity of variances was performed over each stimulus, again excluding Control. Bartlett's test rejected the null hypothesis that all distributions of completion times have the same variance ($\chi^2 = 5521.543,\ p < 0.0001$). These results assert that different stimuli influence subject task performance differently. This suggests that there are different aspects to the specific

stimulus that can be altered to impact performance differently. Implications are discussed in the next section.

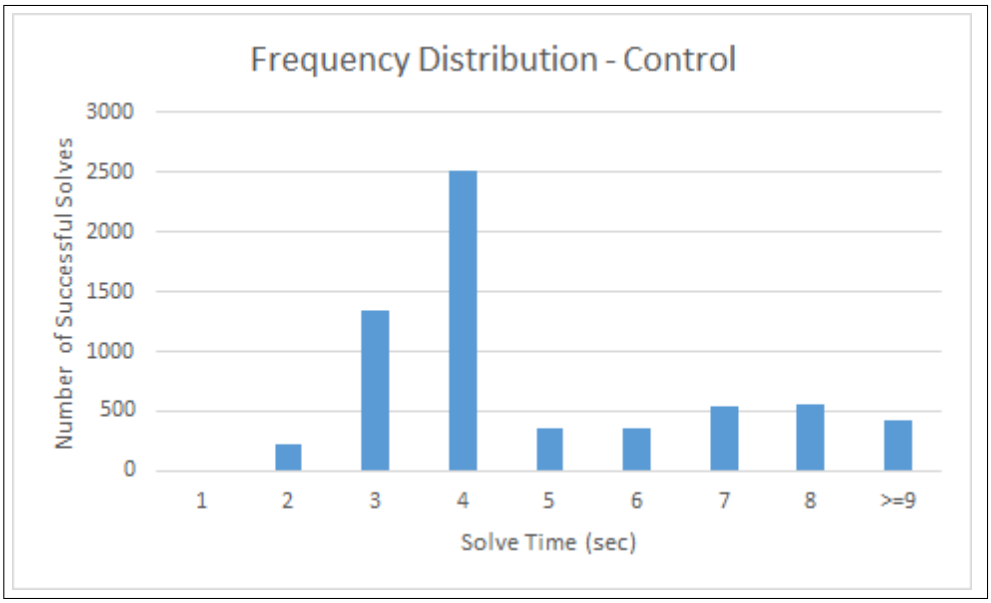


Figure 4.4: Frequency Distribution of Successful Solve Times: Control

Figures 4.4-4.9 show frequency distributions of response times by stimulus. They are similar to exponentially modified Gaussian distributions, consistent with reaction time distributions [64]. This is somewhat expected, since subjects were instructed to solve CAPTCHAs as quickly and as accurately as they could. Although this correlation can help future studies into the cognitive task of completing text-based CAPTCHAs, it is out of the scope of this chapter.

We note that the stimuli with the greatest impact on subject completion times have much heavier tails than other distributions. These correspond to the highly

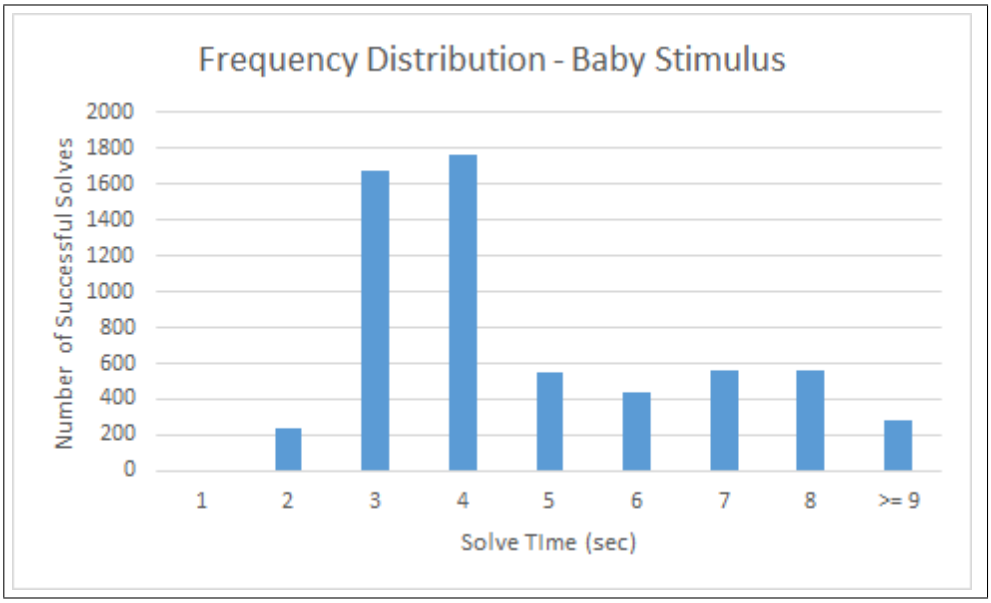


Figure 4.5: Frequency Distribution of Successful Solve Times: Baby Stimulus

dynamic stimuli which also negatively impact subject failure rates. In particular, voice stands out because it is a task-specific stimulus; its exaggerated effect on subject performance is discussed below.

## 4.4 Discussion of Observed Effects

As results show, subjects solving CAPTCHAs are not uniformly impacted by different stimuli. We observed both positive and negative effects. More dynamic or task-specific stimuli f (such as looming, voice and natural) negatively impact subject performance, while the simplest static stimulus (brook) had a positive effect.

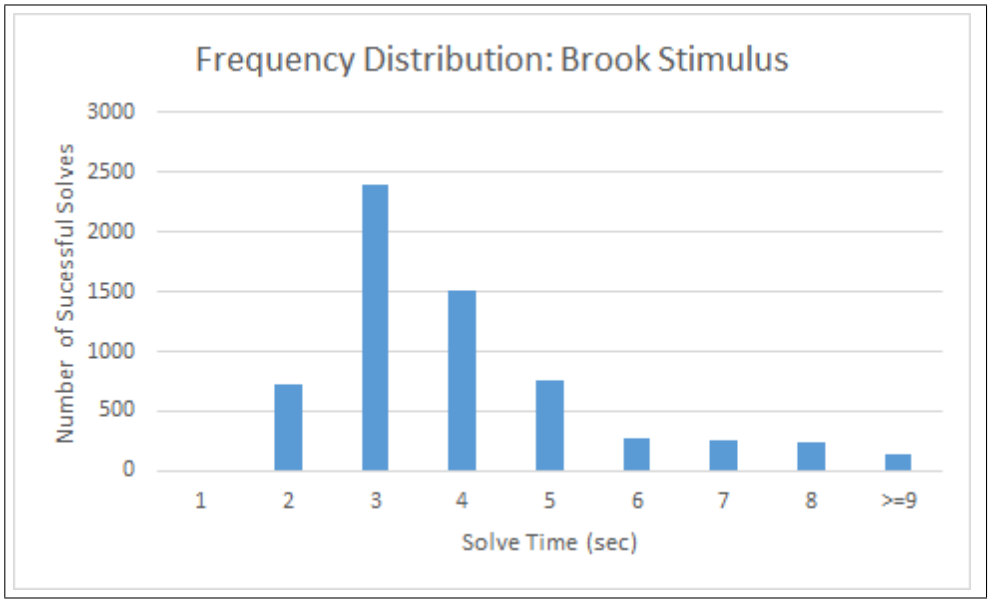


Figure 4.6: Frequency Distribution of Successful Solve Times: Brook Stimulus

Interestingly, crying baby had a substantial negative effect on subject failure rates, though it did not significantly influence subject completion times.

The above is mostly consistent with the Yerkes-Dodson Law, which, states that a subject's overall level of sensory arousal is a determining factor in their performance at any task. At a low level of arousal, a subject is uninterested, and unengaged with the task at hand, and thus does not perform optimally. Similarly, an overstimulated subject is likely to have attention split between the arousing stimuli and the task at hand; thus performance suffers. However, there is a middle ground where a subject's overall arousal level allows being engaged with, yet not overwhelmed by, the task, thus yielding optimal performance. This relationship

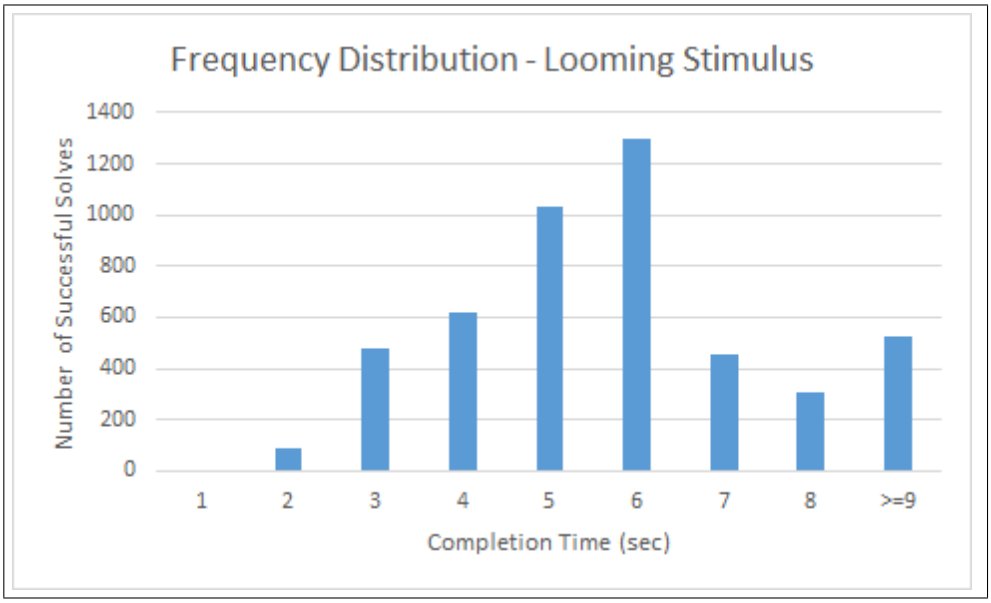


Figure 4.7: Frequency Distribution of Successful Solve Times: Looming Stimulus

between sensory arousal and performance generally follows an upside-down U-shaped curve, as in Figure 5.2 [16]. With this performance curve in mind, we separate further discussion into implications of observed beneficial and negative effects.

### 4.4.1 Beneficial Effects

Only the babbling brook stimulus had a positive impact on subject failure rates and completion times.

It is intuitively obvious that our subjects were not highly engaged with their assigned task. Their general level of sensory arousal in our experiment is sim-

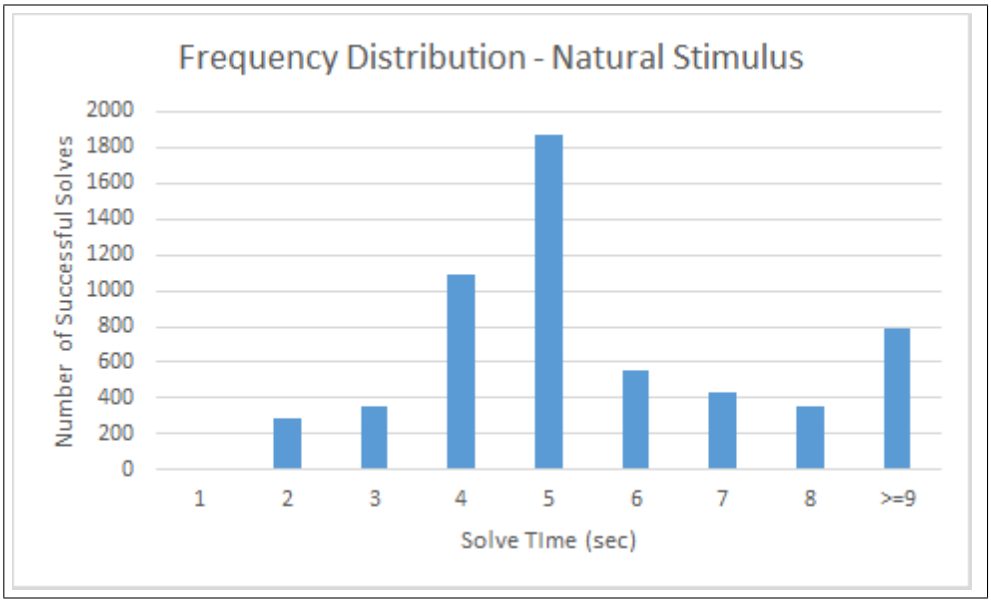


Figure 4.8: Frequency Distribution of Successful Solve Times: Natural Stimulus

ilar to that of one forced to perform a boring and routine security-critical task. Because of this low level of initial engagement, the Yerkes-Dodson Law implies that introduction of additional stimulation can improve task performance. In our case, this resulted in increased speed of correct CAPTCHA completion under the babbling brook stimulus. This simple and static (yet relaxing) stimulus served to pique subject arousal without overwhelming their attentional resources.

The above illustrates the fine line between optimal sensory arousal and over-stimulation. While our subjects might not have been sufficiently engaged with the task at hand, results imply that cognitive resources required to successfully solve CAPTCHAs as quickly as possible left little additional room for stimulation

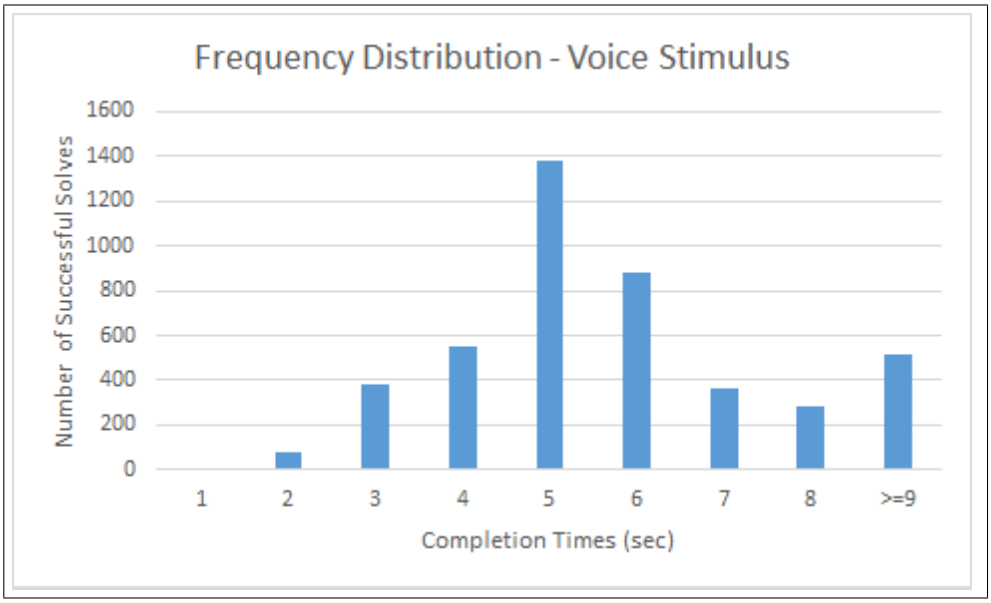


Figure 4.9: Frequency Distribution of Successful Solve Times: Voice Stimulus

before the subject became overstimulated. However, this beneficial effect suggests that there must be a range of stimulation that can reliably improve performance. Thus, there could be a way for benign actors to incorporate sensory stimulation into security-critical tasks (such as CAPTCHAs) to push subjects along the Yerkes-Dodson curve towards a more beneficial level of sensory arousal, yielding better performance.

### 4.4.2 Negative Effects

Several types of auditory stimuli negatively impacted subjects' successful completion. However, collected data shows that this impact is not consistent across

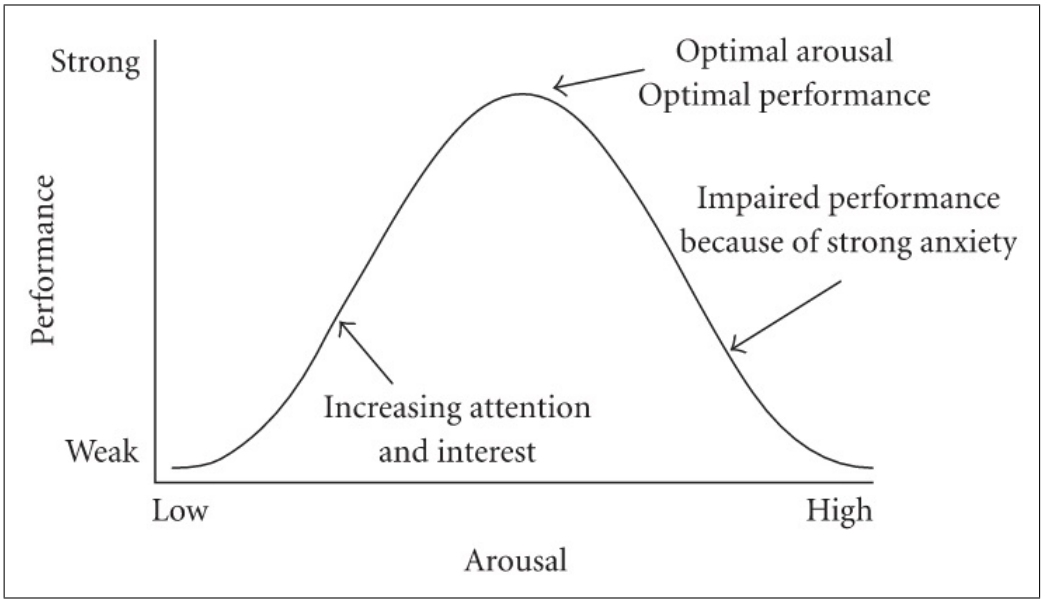


Figure 4.10: The Yerkes-Dodson Relationship Between Sensory Arousal Levels and Performance

all stimuli. The negative effect may be tied to certain features of a particular stimulus. Instances of significant degradation in subject success rates were linked to dynamic sound stimuli, more than static ones. However, this comes with the noted exception of crying baby. While static, it had by far the greatest negative impact on subject failure rates. This could be related to the ecological significance of the sound of a crying baby. In turn, it might be that highly dynamic or aversive stimuli (e.g., Natural or Looming) are not necessarily the most effective adversarial stimuli, despite what the Yerkes-Dodson model asserts. Instead, ecologically-significant stimuli such as crying baby could be crafted for a specific

victim population.

Negative impact on subject task completion rates under these conditions could pave the way for the adversary who controls the ambient soundscape. Through the use of specifically-crafted sounds with shifting intensity levels (or high ecological significance), the adversary could force a user into failing CAPTCHAs as a denial-of-service (DoS) attack. Moreover, not being limited by any ethical boundaries, the adversary can increase the volume far beyond OSHA-recommended safe levels. This would allow creation of even more dynamic stimuli and could push performance degradation beyond the doubling of errors we observed with the crying baby stimulus.

Also, more dynamic stimuli impacted completion speed of successful subjects, slowing them down. The one-way ANOVA analysis we performed on stimuli distributions implies that different stimuli impact completion speeds differently. Furthermore, voice was the stimulus with the greatest impact on subject completion times. This is noteworthy because the task itself revolves around visual interpretation of letters and numbers.

It is reasonable to assume that the subjects are confounded by the sensory crossfire of listening to random letters and numbers being read aloud while they try to read and write random letters and numbers. This is analogous to the Stroop

effect, and suggests that there are some features of the specific stimuli that can impact completion speeds differently [40]. The adversary can use knowledge of the specific task to construct an optimal interfering stimulus.

The real threat of negative effects occurs when they are combined. CAPTCHAs are often used as a defense against the abuse of bots at point-of-sale of limited-quantity time-sensitive services, such as event tickets or travel flash sales. These limited commodities typically sell out completely, within seconds of availability [31]. Therefore, even a single CAPTCHA failure or a second-long delay, can cause a victim to totally miss out on a potentially important (to them) opportunity.

## 4.5 Unattended Setup Analysis

We now discuss advantages and disadvantages of the unattended setup.

### 4.5.1 Advantages

The primary goal of our study was **not** to assess accuracy of the unattended experimental setup. However, results from the control case are analogous to the attended experiment in [14] which used short alphanumeric CAPTCHAs with 1-px. global lines. Results obtained in the control case for our experiment: mean solving time

of $4.62$ seconds and accuracy of $0.912$ for a $5$ character code are consistent with predictions in [14] for the same type of CAPTCHAs. This reinforces equivalence between unattended and attended experimental paradigms.

In general, unattended setups are very well-suited for completing rote, repetitive tasks, such as solving numerous CAPTCHAs. Since subject performance appears to be in-line in both paradigms, an unattended setup saves person-hours that are otherwise spent on logistics of scheduling and physically attending experiments. Moreover, there is no burden on the subject to adhere to a particular schedule, or a limited time-window, since the experiment can run 24/7/365. Furthermore, although it was not done in this case, the unattended paradigm allows for seamless, identical replication in multiple locations simultaneously, which is impossible in an attended manner. Finally, this paradigm entirely avoids experimenter bias: since no one is present during the experiment, there is no way to taint data collection by experimenter's actions.

### 4.5.2 Limitations

As mentioned in the discussion of Data Cleaning, some subjects were non-compliant and their data was discarded. This occurred despite clear instructions (during the initial phase) that CAPTCHAs had to be solved continuously for $54$ minutes. Non-

compliance is a basic limitation of the unattended setup: no one can enforce the rules in real-time[3].

Our setup did not capture fine-grained data about subjects' awareness of the stimuli. In the video recordings of some subjects, there is some evidence of them noticing the stimuli in obvious ways, such as making verbal remarks, or turning their heads towards the speakers. However, there is no firm evidence that shows any subject's failure to notice a given stimulus. Such information would be crucial for development of a realistic adversarial model.

The unattended setup might be both appropriate and useful for assessment of task performance, completion of questionnaires or any study that has subjects act in a fixed manner. However, it is not well-suited for adaptive data collection, e.g., what may be obtained in a loosely-structured interview. Also, since there is no on-site real-time interaction, every subject has an identical experience, which can cause the loss of corner-case data.

## 4.6 Study Shortcomings

This section discusses some shortcomings of the study.

[3]Although it would have been possible to detect non-compliance automatically, e.g., via an inactivity timeout, non-compliant subject data would still be discarded

### 4.6.1 Homogeneous Subjects

Our subject group was comprised of young and tech-savvy college students. This is a consequence of the experiment's location and recruitment methods. Replication of this experiment in a non-academic setting would be useful. However, recruiting an appropriately diverse set of subjects is still difficult, even in a public setting. Ideal venues might be stadiums, concert halls, fairgrounds or shopping malls. Unfortunately, deployment of the unattended setup in such public locations is logistically infeasible. Since such public areas are already full of other sensory stimuli, reliable adjustment of subjects' arousal level in a consistent manner would be very hard. Furthermore, it would be very difficult to secure expensive experimental equipment.

### 4.6.2 Synthetic Environment

Even though we attempted to provide a realistic environment for CAPTCHAs, our setup was obviously a contrived, artificial and controlled space. Typically, people encounter CAPTCHAs while using their own devices from their own homes or offices. As such, it would be intuitive to conduct a study remotely over the Internet. However, this would introduce many compounding and potentially dangerous variables. First, there would be no way of knowing ahead of time the exact

nature of the subjects' auditory environment. This could lead to complications ranging from the trivial nullification of collected data (e.g., if subject's audio-out is muted) all the way to damaging subject's auditory faculties (e.g., in-ear headphones turned to a dangerously high volume).

This further complicates measurement of any effects of auditory stimuli, as it becomes unclear if any two subjects encounter the stimuli the same way. For example, a subject using headphones at a high volume is going to have a drastically different experience than a subject using speakers at a low volume. These differences will confound the actual impact of the stimuli, making it extremely difficult to quantify any meaningful effect on task performance. Because of the need of homogeneity in presentation of the stimuli, it is easy to see how such an online experiment would be ineffective in practice.

## 4.7 Ethical Consideration

Experiments described in this chapter were fully authorized by the Institutional Review Board (IRB) of the university, well before the study. The level of review was: Exempt, Category II. Further IRB-related details are available upon request. No sensitive data was harvested during the experiments and minimal identifying

information was retained. In particular:

1. No names, addresses, phone numbers or other identifying information was collected from the participants.
2. Although email addresses were solicited in order to confirm participation, they were erased very soon thereafter.
3. Video recordings of the experiments were kept for study integrity purposes. However, they were erased before the IRB expiration time.

Finally, with regard to safety, sound levels were maintained at between 70 and 88 dB, which is (especially, for only 2:15 minutes) generally considered safe, as discussed earlier in Section 5.2.

## 4.8 Conclusions

As IoT-enabled sensory environments become more common, the threat of having to complete security-critical tasks in an adversary-controlled environment increases. This trend motivates studying the impact of external stimuli on performance of such tasks. Research described in this chapter sheds some light on the impact of sensory stimulation on performance of security-critical tasks. However, there remain numerous outstanding issues and directions for future work:.

Our results in the context of CAPTCHA highlights the threat of a realistic distributed adversary that aims to induce extra errors and/or longer task completion. While this may not be seen as dire, due to the nature of CAPTCHAs, it opens up a worrisome attack vector for cognitively similar tasks. Notably, many systems implementing two-factor authentication use a similar challenge format to CAPTCHAs, with the distinction that challenges are sent to the user in plain text, instead of a distorted image.

# Chapter 5

# Exploring Effects of Auditory Stimuli on String Entry Tasks

## 5.1 Introduction

Secure, correct and efficient user authentication is an integral component of any meaningful security system. Authentication schemes in the typical modern workplace typically leverages two factors: (1) the user demonstrates knowledge of a secret password or PIN, and (2) the user proves possession of a secure device or token [55]. This second factor seeks to avoid many of the problems associated with knowledge-based authentication by removing the burden of relying on a human

to recall a complex string. Instead,these protocols rely on using a secure hardware token or trusted smartphone application to generate a short-lived key that the user enters alongside their PIN or password [4]. This has led to relatively high adoption rates of smartphone applications such as DUO Mobile [24] and physical tokens such as the RSA Securid token [2].

Despite the variety of protocols, tokens, and applications designed to act as a second factor for authentication, almost no attention has been paid to the user's physical environment while preforming a two-factor authentication task. These tasks are typically performed in noisy environments, such as shared offices, where users can be exposed to a wide variety of sensory stimulation that is outside of the users' explicit control. These stimuli can be either incidental or intentional (i.e., a natural product of the user's surroundings or explicitly introduced by some actor,) and benign or malicious. The impact of these stimuli on the performance of these tasks is unknown, and has not been explored.

This has become especially worrisome as the smart-home environment has emerged and increased in popularity over the last decade. As these sensory environments become more commonplace, it becomes more appealing for adversaries to try to compromise them. A typical smart home represents a veritable buffet of targets for attacks seeking to compromise a victim's physical environment to in-

terfere with their authentication attempts. This is particularly dangerous, for two-factor authentication that are time-sensitive and only allow for a limited number of invalid entries before the victim's account is locked. Once the account is locked, the victim must preform a lengthy recovery process that at best is an additional burden and represents a short-lived denial-of-service, or at worst is a vulnerable process the adversary can exploit to get ownership of the victim's account.

In order to further explore the potential impact of attacks leveraging the compromise of the victim's physical sensory environment, we utilized an unattended experimental environment similar to the setting used in the CAPTCHA study in Chapter 4 to evaluate the performance of subjects attempting a timed short-authentication-string entry task while they are exposed to a variety of unexpected auditory stimuli. We evaluated $53$ subjects in our fully-unattended experimental setting. In line with our previous experiments, we expected highly dynamic stimuli to have the greatest negative impact on subject task performance, and for there to be a positive effect when subjects were exposed to the simplest stimuli. This experiment subverted our expectations, especially with regards to subject completion times in successful cases. We observed no significant departure from the control with any stimulus in subject successful completion times.

The rest of this chapter is organized as follows: The next section outlines

the design and setup of our experiments, followed by experimental results. Next, implications of the results and advantages of the unattended experimental environment are discussed. The chapter concludes with limitations of our approach and ethical considerations.

## 5.2 Methodology

This section describes our experimental setup, procedures and subject parameters.

### 5.2.1 Apparatus

Our experimental setting was designed to allow for fully automated experiments with a wide range of sensory inputs. In a manner similar to our previous work on CAPTCHA entry, we located the experiment in a dedicated office in the Psychology Department building of a large public university. The setup is comprised entirely of the following popular commercial-off-the-shelf (COTS) components:

- Commodity Windows desktop computer with keyboard and mouse
- 2 19" Dell 1907FPc monitors placed side-by-side.
- Logitech C920 HD Webcam.
- Logitech Z200 Stereo Speaker System[1].

[1] with the volume knob physically disabled.

This experimental setup is supposed to mimic the typical workplace environment where a user would log in. Typically, a second device is used for the second factor, instead of a second screen. Due to physical security concerns, we were unable to include a small portable device to serve as the secondary device.

### 5.2.2 Procedures

The experimental environment was entirely unattended. As in our previous experiments evaluating CAPTCHA performance, an instructional PowerPoint presentation was used for subject instruction, instead of a live experimenter. This presentation was each subject's only source of information about the experiment. Actual experimenter involvement was limited to off-line activities: (1) periodic recalibration of auditory stimuli, and (2) occasional repair or repositioning of some components that suffered minor damage or were moved throughout the study's lifetime. This unattended setup allowed the experiment to run without interruption 24/7/365. It was conducted over a 3-month period in the Spring of 2018.

The goal of the experiment was to measure the performance of subjects attempting to correctly respond to as many short-authentication-string entry challenges as possible within a fixed time frame. Subjects were expected to solve them continuously for $54$ minutes. During this period, a subject was exposed to 4

rounds of 6 auditory stimuli. The control and stimuli were presented in a random order within each round, to mitigate any ordering effects on subject performance.

We picked short-authentication-string entry as the security-critical task for several reasons. First, this task does not require the subjects to enter any personally identifying information (PII) or secrets in order to solve it, and can be dynamically generated on the fly, allowing for the study of subject behavior across tens of thousands of solution attempts. This is in contrast with other security-critical tasks, such as password entry, which require PII and demonstrate a clear training effect as the same password is entered repeatedly over the course of an experiment. Second, two-factor authentication is a common task in the modern workplace, and it is reasonable to assume most subjects are familiar with it, as opposed to infrequent tasks such as Bluetooth pairing, which are only preformed once per device-pair. Third, the cognitive effort needed to solve short-authentication-string entry (recognize-and-type) is higher than the simple comparison task in Bluetooth pairing, and is similar to CAPTCHA entry and recall-and-type tasks, such as password entry [54]. Finally, this task is commonly know to be security-critical. Multiple failures on a two-factor authentication task have negative ramifications, such as temporarily losing account access.

The experiment runs in four phases:

1. **Initial:** subject enters the office, sits down at a desktop computer and starts the instructional PowerPoint presentation. Duration: Negligible.
2. **Instruction:** subject is instructed in the nature of short-authentication-strings and the experimental procedure. Duration: 2-4 minutes
3. **Challenge Presentation**: subject is presented with a random short-authentication-string challenge. Upon submitting a solution, a new authentication challenge is presented, regardless of the accuracy of the response.Additionally, if the subject go more than $30$ seconds without providing a response, a new challenge is generated. Subjects are exposed to the stimulus conditions for $24$ rounds, each round lasts $2{:}15$. Duration: $54$ minutes.
4. **Final:** subject is taken to a survey page and asked to enter basic demographic information. Duration: 2-3 minutes

The entire experiment lasts between $58$ and $61$ minutes. Each subject's participation is recorded by the webcam and by screen-capturing software, to ensure compliance with the procedure. Since our objective is to assess overall impact of auditory stimuli on subject performance (and not performance degradation due to a surprise), the first $15$ seconds of each stimulus condition were not used in data collection. This should accurately capture the enduring effect of the auditory stimuli, and ignore the spiking effect (i.e., surprise) on the attentional system due

to the introduction of an unexpected stimulus [54].

### 5.2.3 Stimuli Selection

The experiment consisted of two categories of auditory stimuli: (1) static with single volume level, and (2) dynamic, that changed volume throughout presentation.

Static sound stimuli were the sounds of: (1) crying baby, (2) babbling brook, and (3) human voice reading individual digits in random order at a rate of two per second. (1) was chosen for its biological significance as a source that needs immediate, specific attention, (2) was selected and a relaxing sound typically used in ”white noise’ machines to induce sleep. Stimulus (3) was chosen to interfere with the task-specific cognitive processes used to read and recall numbers. This is analogous to the Stroop effect, a phenomenon wherein subjects who attempt to read the written name of a color that is rendered in a different color (e.g., the word ”red” written in blue ink) do so slower and in a more error-prone way than reading the same words in plain black ink [40]. Specific volumes of the three static stimuli were:

1. Crying baby: 78 dB
2. Babbling brook: 70 dB
3. Human voice: 75 dB

The two dynamic stimuli included: (1) randomly generated looming sounds, and (2) randomly ordered menagerie of natural, aversive sounds. The looming stimulus was an amplitude modulated tone that increased from nearly silent to 85 dB over 5 seconds. Its intensity curve is shown in Figure 5.1. Once the looming sound completed, it repeats at a different Left/Right speaker balance, selected randomly. This repeats continuously for the entire 2:15 minute stimulus window. The natural stimulus consisted of a randomly generated sequence of aversive sounds, which included: circular saw cutting concrete, blaring vuvuzela, nails on a chalkboard, and spinning helicopter rotors. These sounds were played at a randomly selected volume from 75 to 88 dB. Each lasted for up to 2 seconds before changing to the next random sound.

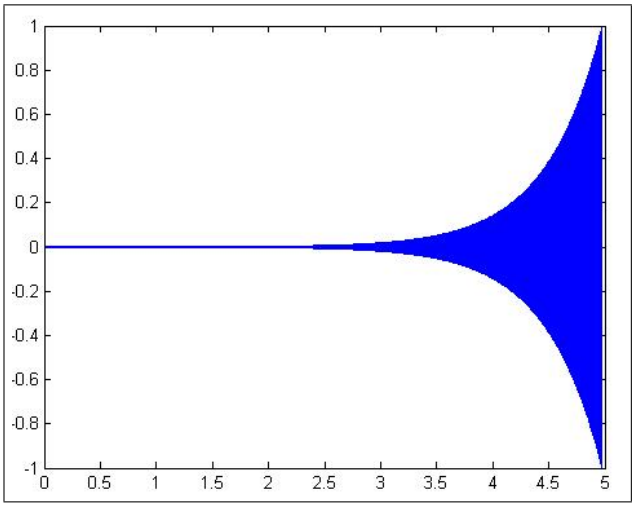


Figure 5.1: Looming Sound Intensity Function

Even the highest stimuli volume (88 dB) is well within the *safe range*, as defined by US Occupational Safety & Health Administration (OSHA) guidelines.[2] This is not a realsitic limitation for an adversary who has no ethical qualms about permanently damaging thier victim's auditory faculties, but such an adversary is out of our scope for this experiment.

### 5.2.4 Initial Hypotheses

Informed by previous experiments, our initial hypothesis was that introduction of unexpected auditory stimuli while responding to string-entry challenges would have a negative impact on subject performance. Particularly with highly dynamic stimuli, we expected two outcomes, as compared to a distraction-free (Control) setting:

**[H1]:** Higher error rates, and

**[H2]:** Longer completion times in successful cases

We hypothesized this because, although mixed results were observed in our previous studies on Bluetooth pairing, string entry is a more difficult cognitive task, similar to CAPTCHA entry, and will use up more of the subjects' attentional re-

[2]OSHA requires all employers to implement a Hearing Conservation Program where workers are exposed to a time-weighted average noise level of 90 dB or higher over an 8 hour work shift. Our noise levels were for a much lower duration, and only the very loudest was within the regulated range. See: `https://www.osha.gov/SLTC/noisehearingconservation/`.

sources even in the distraction-free case [56].

### 5.2.5 Recruitment

Recruitment was handled through the human subjects lab pool of Psychology Department at a large public university. A brief description of the study was posted on an online bulletin, and undergraduate students were allowed to sign up for the experiment and were compensated with course credit. This led to a subject population that reflected the undergraduate population of the university's college of Arts and Sciences; dominated by college-age ($18 - 25$) individuals with a larger female population ($35$ participants, $69\%$) than male ($18$ participants, $31\%$.)

## 5.3 Results

This section discusses the results, starting with data cleaning and proceeding to subject task completion effects.

### 5.3.1 Data Cleaning

A total of $57$ subjects took part in the study. However, $4$ of them were non-compliant with the experimental procedure, and prematurely quit the experiment.

Table 5.1: Subject Failure Rates

| Stimulus | #Successful Entries | #Unsuccessful Entries | Failure Rate | Odds Ratio wrt Control | $p$ |
|---|---|---|---|---|---|
| None (Control) | 17955 | 432 | 0.023 | - | - |
| Baby | 14023 | 735 | 0.050 | 2.17 | $< 0.001$ |
| Brook | 18239 | 388 | 0.021 | 0.91 | 0.084 |
| Looming | 16432 | 592 | 0.027 | 1.52 | $< 0.001$ |
| Natural | 15345 | 493 | 0.041 | 1.17 | $< 0.001$ |
| Voice | 14683 | 621 | 0.029 | 1.78 | $< 0.001$ |
| **Total** | 96677 | 3261 | 0.029 | - | - |

Since this behavior was captured by the recording software, all data from these subjects was discarded.

### 5.3.2 Task Failure Rate

Table 5.1 shows subject failure rates by each stimulus. With the exception of the babbling brook, every stimulus had a statistically significant impact on subject error rates. In each of these cases, the introduction of auditory stimulation served to make subjects more likely to commit errors. In the most extreme case, that of the crying baby stimulus, subjects were over two times more likely to fail the task. The second most impactful stimulus was the task-specific voice stimulus, which

increased failure rates by a factor of $1.78$ over the control. This is of particular interest because these two stimuli are static, and should not create the greatest levels of sensory arousal as per the Brain Arousal Model [56]. Instead, the stimuli which had the greatest impact were those that carried biological or task-specific connotations. The full implications of this increase in subject failure rates, the particulars of the stimuli that caused them, and the corresponding attack space that is made available to an adversary with control over the sensory environment is addressed in detail in Section 5.4

### 5.3.3 Task Completion Times

Table 5.2: Avg Times (sec) for Successful Solutions

| Stimulus | Mean Time | Standard Deviation | DF wrt Control | t-value wrt Control | $p$ | Cohen's D |
|---|---|---|---|---|---|---|
| None (Control) | 2.458 | 1.463 | - | - | - | - |
| Baby | 2.720 | 2.184 | 31976 | 0.103 | 0.918 | 0.14 |
| Brook | 2.257 | 1.100 | 36192 | 15.026 | $0.110$ | 0.08 |
| Looming | 2.092 | 1.929 | 34385 | 17.373 | $0.153$ | 0.21 |
| Natural | 2.409 | 1.770 | 33298 | 18.505 | $0.0255$ | 0.030 |
| Voice | 2.680 | 1.825 | 32636 | 0.0961 | $0.926$ | 0.13 |

Table 5.2 shows average completion times for successful response comple-

tions under each stimulus. Counter-intuitively, our results indicate that none of the stimuli have a significant impact on subject completion times. One possible explanation for this neutral result when compared to previous work on CAPTCHA entry lies in the relative cognitive load of the two tasks. In general, reading and recording a string of transformed letters and numbers, as required by CAPTCHA entry, is more demanding that reading and recording an unaltered string of just numbers. This lower cognitive load could have left subjects with more attentional resource to devote to the classification of the auditory stimuli, giving them a level of resiliency against performance degredations. The implications of these results are discussed in the following section.

## 5.4 Discussion of Observed Effects

As the results show, subjects respond to short-string authentication challenges are not uniformly impacted by exposure to different auditory stimuli. In fact, a wide range of neutral and negative effects were observed. Surprisingly, there was not a direct correlation between the highly dynamic stimuli and large negative impacts on subject performance, although the simplest stimulus, the babbling brook, did not have a significant effect on any aspect of subject performance.

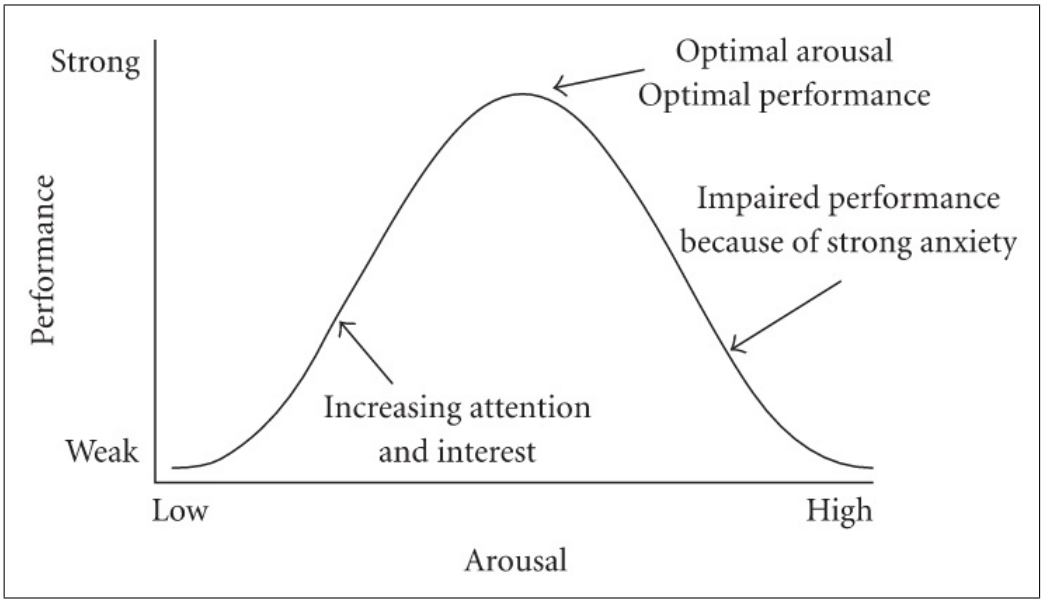


Figure 5.2: The Yerkes-Dodson Relationship Between Sensory Arousal Levels and Performance

The above is partially consistent with the Yerkes-Dodson Law, which states that a subject's overall level of sensory arousal is the chief determinant of task performance. In brief, subjects who are at a low level of sensory arousal are not engaged with the task at hand, and can be error prone. On the other end of the spectrum, an overly high level of sensory arousal can overwhelm a subject, forcing them to multitask and become error-prone.This relationship between sensory arousal and performance generally follows an upside-down U-shaped curve, as in Figure 5.2 [16]. However, the stark negative impact of the relatively static crying baby and voice stimuli imply that there is a greater space to interfere with subject

performance than just increasing levels of sensory arousal. With these differing impact on subject performance in mind, we separate our discussion by negative effects, and neutral effects.

### 5.4.1 Negative Effects

The majority of the auditory stimuli negatively impact subject completion rates when responding to the short-string challenges. However, this impact is not uniform across the stimuli, nor is it directly proportional to the dynanism of the stimuli, as our previous experiments would incline us to believe. In particular, the crying baby stimulus, which is a relatively static stimulus that the attentional system should be able to quickly classify and disregard, was shown to have the greatest negative impact.

While we lack a sufficient sample size between males and females to make statistically significant claims, it is interesting to note that in our earliest experiments, which were male-dominated, the crying baby stimulus did not have a unique impact on either subject completion times or failure rates, while in our later experiments, which were female-dominated, the crying baby stimulus had a markedly greater negative impact on subject failure rates than any other stimulus. Given the demographics of our subject populations, it is not unreasonable

to assume that there is some biological trigger that a crying baby activates that is specific to females in the 18-29 age group that may not be shared with males of the same age. This serves to highlight the danger of an adversary who is targeting a specific individual or demographic groups of individuals. The more biologically or culturally-relevant information that can be gathered about a potential victim, the more easily a targeted stimulus could be crafted that, on its face, does not appear to be explicitly malicious or aversive.

Similarly, the second largest effect size was observed with the task-specific voice stimulus. It is reasonable to assume that the subjects are confounded by the sensory crossfire of listening to random numbers being read aloud while they try to read and write seemingly random numbers. This is analogous to the Stroop effect, and suggests that there are some features of the specific stimuli that can impact task performance without being overtly aversive [40]. The adversary can use knowledge of the specific task to construct an optimal interfering stimulus.

Finally, as expected, there was a general negative impact on task completion rates caused by highly dynamic stimuli. This is consistent with the Brain Arousal Model, and can serve as a baseline stimulus type for an adversary with control over the sensory environment, but little knowledge of the specific tasks their potential victim would be preforming. Across all experiments we have conducted

for all security-critical tasks, there has been some negative impact cause by exposure to highly dynamic sensory stimuli, and it is reasonable to infer that this is a generalizable result that will hold for arbitrary security-critical tasks.

A greater understanding of the causes of negative impacts on subject task completion rates can enable an adversary who controls a specific victim's sensory environment to target that victim to induce errors. Specifically from the results in this experiment it appears that the inclusion of a targeted biologically-relevant or task-specific stimulus can maximize the potential disruption an adversary is able to cause. By forcing the victim into failing authentication requests, the adversary could trigger a lockout on the victim's account as part of a two-phase attack. In the least damaging case, this represents additional user burden, and serve as a denial-of-service attack. At worst, once the account is locked, the adversary could then conduct a man-in-the-middle attack on the more vulnerable account recovery process and gain access to the victim's account.

### 5.4.2 Neutral Effects

Interestingly, there were no statistically significant effects, positive or negative, on subject completion times for successful responses. While this may not intuitively present itself as an important result, the implications of the existence of

stimuli that an adversary can use to ”target” one aspect of their victim’s experience are worrisome. This could provide an adversary with a less-noticeable custom-tailored attack wherein the victim may not be aware of ephemeral sensory stimulation that only lasts long enough for the victim to fail their task.

This serves to highlight the dangers of an adversary who is intimately familiar with the victim as well as the task they are trying to complete. A highly knowledgeable adversary is more likely to exploit their knowledge of the space of sensory stimulation to custom-craft a stimulus that is not inherently aversive, or even ”annoying” while still being able to target the exact aspect of performance they wish to degrade.

## 5.5 Unattended Setup Analysis

We now discuss advantages and disadvantages of the unattended setup.

### 5.5.1 Advantages

In general, the unattended experimental paragdim has proven itself invaluable for the completion of rote, repetitive tasks, such as responding to two-factor authentication challenges hundreds of times. The unattended setup serves to save person-

hours that would otherwise be spent enforcing compliance rules that, as discussed in Section 3.3.1, are violated very infrequently in practice, and are trivially detectable when violations occur. Furthermore, experimenter involvement is a well known source of potential bias and compromise of collected data. Also, subjects are not forced to adhere themselves to an experimenter's schedule, allowing the experiment to run 24/7/365. Finally, this paradigm allows for the rapid redeployment or simultaneous deployment of identical experiments across multiple locations without compromising the homogeneity of the subject experience..

### 5.5.2 Limitations

As mentioned earlier in Section 5.3.1, there was a number of non-compliant subjects which had to be discarded. This happened even though the instructional PowerPoint clearly stating that subjects were to respond as quickly as possible to the challenges for the entire time that they are being presented. Non-compliance is an inherent limitation of any unattended setup; since no experimenter is physically present, real-time enforcement of the experiment's rules was not possible, even though such abuses are trivially detectable in post-factum review.

Our experimental environment was unable to capture fine-grained data about subjects' implicit awareness of the stimuli. In review of the video recordings of the

subjects, it was occasionally the case that a subject would either visually or verbally remark on some of the stimuli as they were being presented, we had no way of verifying a subjects' failure to notice a given stimulus. Even though this data would be greatly beneficial for the construction of an exact adversarial model, the precise collection of a subject's cognitive response to a stimulus requires physical instrumentation of the subject (e.g., using an electroencephalogram (EEG) headset) and would be impossible in any unattended context.

## 5.6 Study Shortcomings

This section discusses some shortcomings of the study.

### 5.6.1 Homogeneous Subjects

As was expected from the experiment's location, as well as our recruitment methods, the vast majority of our subject population were young and tech-savvy college students. The unattended nature of this experiment naturally fits replication, and conducting a follow-on in a non-academic setting would be beneficial to understand the general effects of auditory stimulation. However, recruiting a truly diverse, representative set of subjects would still be logistically difficult, especially

in a public setting. Since public settings are already replete with sensory stimuli that cannot be controlled, it would be impossible to ensure a uniform experience across subjects. That would make evaluating whether changes in performance are the result of our crafted stimuli, or just a product of the environment. Finally, in a truly public setting, securing expense experimental equipment against theft or damage would be difficult.

### 5.6.2 Synthetic Environment

Even though the core advantage of the unattended experimental environment is to provide a more realistic setup than traditional attended experiments conducted in a lab-like setting, our setup was still obviously a contrived, artificial and controlled space. This is most notable in the delivery of the short-string challenges through the use of a secondary screen instead of a small auxiliary device, such as a smartphone or a hardware token. In most real-world deployments of a two-factor authentication scheme using short-string challenges, users log into their own devices, and respond to a challenge generated on their personal auxiliary device. Intuitively, it stands that an experimental design where subjects sign up online to participate using their own devices would serve to approximate the real world as closely as possible. However, such a study would introduce several un-

controllable, and potentially dangerous variables. First, and most importantly, it would be impossible to know the auditory environment beforehand. Participants using hefadphones turned up to maximum volume would undoubtedly have a different experience than participants using speakers set at a low volume, and such data would not be available to the experimenters. These impacts of this range from trivially nullifying any delivered stimuli (e.g., if the audio-out on the participant's personal device is disabled) to the dangerous (e.g., If the participant is using headphones at an unsafe volume for the stimuli.) Secondly, these discrepancies between participants' auditory environments create an untraceable series of knock-on effects. Since the environment is not homogenized, it would become very difficult, if not impossible, to generalize the impact of stimuli across multiple subjects, making such an online experimental design ineffective.

## 5.7 Ethical Consideration

Experiments described in this chapter were fully authorized by the Institutional Review Board (IRB) of the university, well before the study. The level of review was: Exempt, Category II. Further IRB-related details are available upon request. No sensitive data was harvested during the experiments and minimal identifying

information was retained. In particular:

1. No names, addresses, phone numbers or other identifying information was collected from the participants.
2. Although email addresses were solicited in order to confirm participation, they were erased very soon thereafter.
3. Video recordings of the experiments were kept for study integrity purposes. However, they were erased before the IRB expiration time.

Finally, with regard to safety, sound levels were maintained at between 70 and 88 dB, which is (especially, for only 2:15 minutes) generally considered safe, as discussed earlier in Section 5.2.

## 5.8 Conclusions

The wide-ranging proliferation of IoT devices, especially those in the smart home suite, leads to a drastic increase in instrumented sensory environments. These instrumented environments create a plethora of new attack vectors for an adversary who is able to compromise these IoT devices. This trend motivates our study of the impact of external stimuli on the performance of critical tasks. That being said, there are still many unexplored aspects of the cognitive relationship between

task performance and sensory stimulation left for future work.

Our results in the context of Short-String comparison and entry show that the introduction of auditory stimuli has mixed impacts on task failure rates, and negligible impact on overall times in cases of successful completion. While this has not shown performance impacts as stark as those that were shown in the previous chapters of this dissertation, this evaluation serves to further expand our understanding of the full, generalizable, space of security-critical tasks. In particular, our results highlights the potential impact of well-crafted task or target-specific stimuli. An interesting direction to move into in the future would be to evaluate subjects performing several different security-critical tasks while exposed to adversarial auditory noise to evaluate the difference in effect sizes for uniform stimuli across a general space of tasks.

# Chapter 6

# Related Work

This chapter overviews related work in automated experiments, and human-assisted security methods.We also provide background information in psychology, particularly effects of sensory arousal on task performance, as well as effects of visual stimuli on arousal level and emotive state.

## 6.1 Automated Experiments

We are unaware of any prior usability studies utilizing a fully automated and unattended physical environment.

However, some prior work reinforces the validity of virtually-attended remote experiments and unattended online surveys, in contrast with same efforts in a

traditional lab-based setting. Ollesch et al.[48] collected psychometric data in: (1) a physically attended experimental lab setting and (2) its virtually attended remote counterpart. No significant differences between the two sets were found. This is further reinforced by Riva et al. [52] who compared data collected from (1) unattended online, and (2) attended offline, questionnaires. Finally, Lazem and Gracanin [39] replicated two classical social psychology experiments where both the participants and the experimenter were represented by avatars in Second Life[1], instead of being physically co-present. Here too, no significant differences were observed.

## 6.2 User Studies of Secure Device Pairing

Secure device pairing (mostly, but not only, via Bluetooth) has been extensively researched by experts in both security and usability. While initially pairing, the two devices have no prior knowledge of one another, i.e., there is no prior security context. Also, they can not rely on either a Trusted Third Party (TTP) or a Public Key Infrastructure (PKI) to facilitate the protocol. This makes device pairing especially vulnerable to man-in-the-middle (MiTM) attacks. This prompted the design of numerous protocols requiring human involvement (integrity verifi-

[1]See `secondlife.com`

cation) over some out-of-band (OOB) channel, e.g., visual or audio comparison or copying/entering numbers.

For example, Short Authenticated String (SAS) protocols ask the user to compare two strings of about 20 bits each [38].

Uzun et al. [60] performed the first usability study of Bluetooth pairing techniques using SAS. It determined that the "compare-and-confirm" method – which involves the user comparing two 4-to-6-digit decimal numbers and indicating a match or lack thereof – was the most accurate and usable approach.

Kobsa et al. [35] compiled a comprehensive comparative usability study of eleven major secure device pairing methods. They measured task performance times, completion times, completion rates, perceived usability and perceived security. This led to the identification of most problematic as well as most effective pairing methods, for various device configurations.

Goodrich et al. [23] proposed an authentication protocol that used "Mad-Lib" style SAS. Each device in this protocol creates a nonsensical phrase based on the protocol outcome, and the user then determine if the two phrases match. This approach was found to be easier for non-specialist users.

Kainda et al. [29] examined usability of device pairing in a group setting. In this setting, up to 6 users tried to connect their devices to one another by participat-

ing in a SAS protocol. It was found that group effort decreased the expected rate of security and non-security failures. However, if a single individual was shown a SAS different from that of all others participants, the former often lied about the SAS in order to fit in with the group, demonstrating so-called "insecurity of conformity."

Gallego et al. [22] discovered that subject's performance in secure device pairing could be improved if it were to be scored. In other words, notifying subjects about their performance score resulted in fewer errors.

There has been no previous evaluation of the impact of sensory stimulation on the performance of pairing tasks.

## 6.3 User Studies of Text-Based CAPTCHAs

Given ubiquity of CAPTCHAs, it is surprising that only a few usability studies have been conducted.

Chellapilla et al. [15] performed the first usability evaluation of CAPTCHAs, by examining character-based CAPTCHAs and evaluating Robustness/Usability tradeoffs. Results showed that sophisticated segmentation algorithms can violate robustness goals of popular, currently deployed text-based CAPTCHAs. How-

ever, service providers are hesitant to switch to more difficult CAPTCHAs for fear of low user acceptability.

Bursztein et al. [13] conducted a large-scale evaluation of user performance with several CAPTCHA schemes. Performance varied widely from scheme to scheme, with user's success rates ranging from $91\%$ to $70\%$. This contradicted self-reported statistics, e.g., from Ebay, which claimed a $98\%$ successful completion rate. Audio-only CAPTCHAs were found to be extremely difficult for most users, with success rates as low as $35\%$. This motivates guidelines for user-friendly text-based, and the need for further study of audio-only, CAPTCHAs.

Yan and El Ahmed [21] examine what makes CAPTCHAs usable, and non-intrusive. Color is identified as the primary culprit in intrusiveness, as clashing schema can interfere with presentation of the site itself. Furthermore, coloring a CAPTCHA lowers robustness, since it gives an easy target for segmentation, i.e., separating the image by color. Surprisingly, inclusion of color in a CAPTCHA is claimed to be a benefit for both usability and robustness if done correctly. However, what constitutes correct color usage is left as an open problem.

Khalil et al. examine the impact of alphabet familiarity on CAPTCHA performance using different character sets [32]. Familiarity with the alphabet used to construct a text-based CAPTCHA does not impact error rates. However, users'

satisfaction is positively correlated with their familiarity level with the alphabet being used.

Burszstein et al. [14] paramaterized CAPTCHA features to find the most usable combination. This was done with particular focus on low-security CAPTCHAs that could sacrifice robustness and allow bots to achieve $> 0.01\%$ success rate. Subjects were found to prefer CAPTCHAs composed of English-language words with positive connotations (such as "cutest") with simple global distortions, and very few intersection or occluding lines. The study concluded with a candidate CAPTCHA design that showed a $95.4\%$ success rate.

There has been no evaluation of user performance with CAPTCHAs in a noisy environment.

## 6.4 User Studies of Two-Factor Authentication

Many different techniques have been proposed and evaluated as a second factor for user authentication. Authentication techniques fall into one of three types: 1.) What you know (e.g., a password,) 2.) What you own (e.g., token-based authentication,) or 3.) What you are (e.g., biometric authentication). Typically, the first factor in an authentication is password/PIN entry, and falls into the "what you

know" category. While there have been many interesting proposed schemes based on biometric authentication, we did not evaluate the usability of such techniques, as they are fraught with enrollment issues that would make effective large-scale studies in our unattended style infeasible [58].

Instead, we choose to focus on second factors that rely on something you own, in particular protocols that generate a short challenge string for users to enter via a secure hardware token or application on their personal device.

In general, techniques relying on a owned device focus on either secure tokens, such as the RSA Securid token [2] or smartphone applications such as Duo Mobile [24]. Similar to studies of Bluetooth Pairing protocols, it has been found that humans can manage about 20 bits of information [63].

Additionally, there are techniques that leverage a central server to send a one-time PIN to users via email or SMS [41]. While these systems are generally usable, they have a long vulnerability window and reduce in the worst case to the security of the user's email account [43]. Cristofaro et al. conducted a comparative user study of many different two-factor authentication techniques and found that, regardless of context, these tasks are viewed generally favorably, and are considered usable [19]. Many different timing windows have been tested, with a window of 30 seconds or longer being found ideal [11]. The greatest burden was

limited to the physical management of the token [25].

No previous studies have been conducted assessing the impact of sensory stimulation on the performance of these tasks.

## 6.5 Effects of Sensory Stimulation

Sensory stimulation has variable impact on task performance. This is due to many contributing factors, including the subject's current level of arousal. The Yerkes-Dodson Law stipulates an inverse quadratic relationship between arousal and task performance [16]. It implies that, across all contributing stimulants, subjects who are either at a very low, or very high, level of arousal are not likely to perform well, and there exists an optimal level of arousal for correct task completion.

An extension to this law is the notion that completion of less complex tasks that produce lower levels of initial arousal in subjects benefits from inclusion of external stimuli. At the same time, completion of complex tasks that produce a high level of initial arousal suffers from the inclusion of external stimuli. Hockey [27] and Benignus et al. [7] classified this causal relationship by defining the complexity of a task as a function of the task's event rate (i.e., how many subtasks must be completed in a given time-frame) and the number of sources that orig-

inate these subtasks. External stimulation can serve to sharpen the focus of a subject at a low arousal level, improving task performance [49]. Conversely, it can overload subjects that are already at a high level of arousal, and induce errors in task completion [26].

O'Malley and Poplawsky [50] argued that sensory noise affects behavioral selectivity. Specifically, while a consistent positive or negative effect on task completion may not occur, a consistent negative effect was observed for tasks that require subjects to react to signals on their periphery. Meanwhile, a consistent positive effect on task completion was observed for tasks that require subjects to react to signals in the center of their field of attention. This leads the authors to claim that sensory stimulation has the effect of narrowing the subject's area of attention.

In addition to being general external stimuli that serve to raise arousal level, visual stimuli, particularly colors, have social and emotional implications. Naz and Epps [44] surveyed 98 college students about their emotional responses to five principal hues (red, blue, purple, green and yellow), five intermediate hues (yellow-red, green-yellow, blue-green, and red-purple) as well as three achromatic colors (white, gray, and black.) They found that principal hues are more likely to foster positive emotive responses. Furthermore, different colors within each group

induce differing levels of arousal: some (red or green-yellow) increase arousal, while others (blue and green) are perceived as relaxing.

Moreover, visual stimuli were found to be dominating in multi-sensory contexts. Eimer [20] showed that in experiments with tactile, visual, and audio stimuli, subjects overwhelmingly utilized visual queues to localize tactile and auditory events.

# Chapter 7

# Conclusions & Future Work

This dissertation provided usability evaluations for many general security-critical tasks when subjects are exposed to sensory stimulation.

Chapter 3 explored the impacts of auditory stimulation on subjects performing Bluetooth Pairing. It was notable that static auditory stimuli served to improve subject performance wihile the dynamic looming stimulus degraded it. We began to note a relationship between the overall arousal level induced in the subjects and their task performance. From this, we were able to suggest future studies to explore the full space of subject arousal; both to define an attack space for adversaries who own an IoT cyber-physical environment, and for benefactors seeking to create the most pleasant user experience possible. Chapter 4 examined the im-

pact of visual simulation on subjects performing Bluetooth Pairing. Contrary to the previous study on auditory stimuli, there were no positive impacts on subject performance when they were exposed to visual stimuli. However, there was still a correlation between the dynamism of the stimuli and the degree of the negative impact. Even though we explored the effects of both auditory and visual simulation on the performance of Bluetooth Pairing, we have only scratched the surface of the space of stimulating effects. An appealing extension to this work would be to further flesh out impact of different stimuli, with a three-fold intention of:

1. Finding the category stimuli that maximizes subject performance (i.e., the peak of the Yerkes-Dodson curve.)
2. Identifying the point of subject arousal where performance begins to degrade.
3. Classifying a space of stimuli that highly impact subject performance in a negative way, while seeming to be benign.

Chapter 5 evaluated subjects responding to CAPTCHA challenges when exposed to auditory stimuli. In addition to separating stimuli into 'static' and 'dynamic' categories, we also introduced a stimulus that was designed to interfere with subject performance in a task-specific way. Not surprisingly, this new stimulus type had a pronounced negative effect, but it did not have the largest impact

on subject performance. Interestingly, the stimulus with the greatest impact was of biological importance to our subject population. These insights foster interest in a continuation of this study looking at fine-tuning stimuli for both task-specific interference, as well as targeting specific populations.

Finally, in Chapter 6 we reported on subject task performance in answering short-authentication-string entry challenges in a timed, two-factor authentication style under auditory stimulation. This further reinforced our notions of the general relationship between subject performance and stimulation. In this case, however, while there was the expected . Given that the short-authentication-string entry task was considerably less cognitively demanding than the CAPTCHA task conducted in a similar setup, these results may betray the impact of task complexity on subject performance. The natural extension to this is to conduct a study in which subjects perform many differing tasks under adversarial noise in an attempt to identify the base level of sensory arousal for tasks of varying complexity.

All of the studies described in this dissertation were concerned with task performance without any adversarial interference in the presentation of the task itself. The final direction for future work that we intend to pursue is the impact of adversarial noise on the completion of security-critical tasks while subjects are *under attack*. For example, in the case of Bluetooth Pairing, instead of faithfully

showing the correct codes on both the subjects' owned devices as well as our experimental device, we would occasionally show mis-matching codes, representing the case where an adversary is trying to trick the subject into pairing with a malicious device. Any changes in attacker success rate when compared to a traditional noiseless study would further define the advantages offered to the adversary who owns the victim's sensory environment.

# Bibliography


[1] Number of Internet of Things (IoT) devices connected worldwide in 2017 and 2018, by selected type (in millions). https://www.statista.com/statistics/789615/worldwide-connected-iot-devices-by-type/. Accessed: 2018-11-4.

[2] RSA SecurID hardware tokens. https://www.rsa.com/en-us/products/rsa-securid-suite/rsa-securid-access/securid-hardware-tokens. Accessed: 2018-10-30.

[3] AIELLO, J. R., AND DOUTHITT, E. A. Social facilitation from triplett to electronic performance monitoring. *Group Dynamics: Theory, Research, and Practice 5*, 3 (2001), 163–180.

[4] ALOUL, F., ZAHIDI, S., AND EL-HAJJ, W. Two factor authentication using mobile phones. In *Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on* (2009), IEEE, pp. 641–644.

[5] BALFANZ, D., DURFEE, G., SMETTERS, D. K., AND GRINTER, R. E. In search of usable security: Five lessons from the field. *IEEE Security & Privacy 2*, 5 (2004), 19–24.

[6] BECKER, A. Bluetooth security & hacks. Master's thesis, Ruhr-Universität Bochum, Bochum, Germany, 2007.

[7] BENIGNUS, V. A., OTTO, D. A., AND KNELSON, J. H. Effect of low-frequency random noises on performance of a numeric monitoring task. *Perceptual and Motor Skills 40*, 1 (1975), 231–239.

[8] BERG, B. G., KACZMAREK, T., KOBSA, A., AND TSUDIK, G. An exploration of the effects of sensory stimuli on the completion of security tasks. *IEEE Security & Privacy 15*, 6 (2017), 52–60.

[9] BISDIKIAN, C. An overview of the Bluetooth wireless technology. *IEEE Communications Magazine 39*, 12 (2001), 86–94.

[10] BRANT, L. J., AND FOZARD, J. L. Age changes in puretone hearing thresholds in a longitudinal study of normal human aging. *The Journal of the Acoustical Society of America 88*, 2 (1990), 813–820.

[11] BRAZ, C., AND ROBERT, J. M. Security and usability: The case of the user authentication methods. In *Proceedings of the 18th Conference on l'Interaction Homme-Machine* (2006), ACM, pp. 199–203.

[12] BROWN, M. S., AND LITTLE, H. A. Methods and devices for facilitating Bluetooth pairing using a camera as a barcode scanner, Nov. 4 2014. US Patent 8,879,994.

[13] BURSZTEIN, E., BETHARD, S., FABRY, C., MITCHELL, J. C., AND JURAFSKY, D. How good are humans at solving CAPTCHAs? A large scale evaluation. In *Security and Privacy (SP), 2010 IEEE Symposium on* (2010), IEEE, pp. 399–413.

[14] BURSZTEIN, E., MOSCICKI, A., FABRY, C., BETHARD, S., MITCHELL, J. C., AND JURAFSKY, D. Easy does it: More usable CAPTCHAs. In *Proceedings of the 32nd annual ACM Conference on Human Factors in Computing Systems* (2014), ACM, pp. 2637–2646.

[15] CHELLAPILLA, K., LARSON, K., SIMARD, P., AND CZERWINSKI, M. Designing human friendly human interaction proofs (HIPs). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2005), ACM.

[16] COHEN, R. A. Yerkes–Dodson law. In *Encyclopedia of Clinical Neuropsychology*. Springer, 2011, pp. 2737–2738.

[17] CRANOR, L. F. A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology and Security* (2008), USENIX Association, pp. 1–15.

[18] DAVIDOFF, S., LEE, M. K., YIU, C., ZIMMERMAN, J., AND DEY, A. K. Principles of smart home control. In *International Conference on Ubiquitous Computing* (2006), Springer, pp. 19–34.

[19] DE CRISTOFARO, E., DU, H., FREUDIGER, J., AND NORCIE, G. A comparative usability study of two-factor authentication. *arXiv preprint arXiv:1309.5344* (2013).

[20] EIMER, M. Multisensory integration: How visual experience shapes spatial perception. *Current Biology 14*, 3 (2004), R115–R117.

[21] EL AHMAD, A. S., YAN, J., AND NG, W.-Y. CAPTCHA design: Color, usability, and security. *IEEE Internet Computing 16*, 2 (2012), 44–51.

[22] GALLEGO, A., SAXENA, N., AND VORIS, J. Exploring extrinsic motivation for better security: A usability study of scoring-enhanced device pairing. In *Financial Cryptography and Data Security*, A.-R. Sadeghi, Ed., vol. 7859 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 60–68.

[23] GOODRICH, M. T., SIRIVIANOS, M., SOLIS, J., SORIENTE, C., TSUDIK, G., AND UZUN, E. Using audio in secure device pairing. *International Journal of Security and Networks 4*, 1 (2009), 57–68.

[24] GRAVEL, V., GAGNON, F., LECLERC, M., HEMON, M., AND GAGNON, F. Secure authentication system and method, Aug. 11 2011. US Patent App. 13/021,140.

[25] GUNSON, N., MARSHALL, D., MORTON, H., AND JACK, M. User perceptions of security and usability of single-factor and two-factor authentication in automated telephone banking. *Computers & Security 30*, 4 (2011), 208–220.

[26] HARRIS, W. *Stress and Perception: The Effects of Intense Noise Stimulation and Noxious Stimulation upon Perceptual Performance*. Ph.D. thesis, University of Southern California, 1960.

[27] HOCKEY, G. R. J. Effect of loud noise on attentional selectivity. *The Quarterly Journal of Experimental Psychology 22*, 1 (1970), 28–36.

[28] KACZMAREK, T., KOBSA, A., SY, R., AND TSUDIK, G. An unattended study of users performing security critical tasks under adversarial noise. In *Proceedings of the NDSS Workshop on Useable Security 2015*, pp. 14:1–14:12.

[29] KAINDA, R., FLECHAIS, I., AND ROSCOE, A. W. Usability and security of out-of-band channels in secure device pairing protocols. *Proceedings of the 5th Symposium on Usable Privacy and Security* (2009), 11:1–11:12. ACM ID: 1572547.

[30] KAINDA, R., FLECHAIS, I., AND ROSCOE, A. W. Two heads are better than one: Security and usability of device associations in group scenarios. In *Proceedings of the Sixth Symposium on Usable Privacy and Security* (2010), SOUPS '10, pp. 5:1–5:13. ACM ID: 1837117.

[31] KAISER, E., AND FENG, W.-C. Helping ticketmaster: Changing the economics of ticket robots with geographic proof-of-work. In *INFOCOM IEEE Conference on Computer Communications Workshops, 2010* (2010), IEEE, pp. 1–6.

[32] KHALIL, A., ABDALLAH, S., AHMED, S., AND HAJJDIAB, H. Script familiarity and its effect on CAPTCHA usability: An experiment with Arab participants. *International Journal of Web Portals (IJWP) 4*, 2 (2012), 74–87.

[33] KLINE, D. W., AND SCHIEBER, F. Vision and aging. In *Handbook of the Psychology of Aging* (1985), Van Nostrand Reinhold, pp. 296–331.

[34] KOBSA, A., NITHYANAND, R., TSUDIK, G., AND UZUN, E. Can Jannie verify? Usability of display-equipped RFID tags for security purposes. *Journal of Computer Security 21*, 3 (Jan. 2013), 347–370.

[35] KOBSA, A., SONAWALLA, R., TSUDIK, G., UZUN, E., AND WANG, Y. Serial hook-ups: A comparative usability study of secure device pairing methods. *Proceedings of the 5th Symposium on Usable Privacy and Security* (2009), 10:1–10:12. ACM ID: 1572546.

[36] KOELEGA, H. S., BRINKMAN, J. A., ZWEP, B., AND VERBATEN, M. N. Dynamic vs static stimuli in their effect on visual vigilance performance. *Perceptual and Motor Skills 70*, 3 (1990), 823–831.

[37] KOLIAS, C., KAMBOURAKIS, G., STAVROU, A., AND VOAS, J. DDoS in the IoT: Mirai and other botnets. *Computer 50*, 7 (2017), 80–84.

[38] LAUR, S., ASOKAN, N., AND NYBERG, K. Efficient mutual data authentication using manually authenticated strings. Cryptology ePrint Archive, Report 2005/424, 2005. `http://eprint.iacr.org/`.

[39] LAZEM, S., AND GRACANIN, D. Social traps in second life. In *2010 Second International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)* (Mar. 2010), pp. 133–140.

[40] MACLEOD, C. M. Half a century of research on the stroop effect: An integrative review. *Psychological Bulletin 109*, 2 (1991), 163.

[41] MARMIGERE, G., AND SZALAI, Z. System and method for SMS authentication, Nov. 13 2007. US Patent 7,296,156.

[42] MATTERN, F., AND FLOERKEMEIER, C. From the internet of computers to the internet of things. In *From Active Data Management to Event-Based Systems and More*. Springer, 2010, pp. 242–259.

[43] MULLINER, C., BORGAONKAR, R., STEWIN, P., AND SEIFERT, J.-P. SMS-based one-time passwords: Attacks and defense. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment* (2013), Springer, pp. 150–159.

[44] NAZ, K., AND EPPS, H. Relationship between color and emotion: A study of college students. *College Student J 38*, 3 (2004), 396.

[45] NEELON, M. F., WILLIAMS, J. C., AND GARELL, P. C. Elastic attention: Enhanced, then sharpened response to auditory input as attentional load increases. *Frontiers in Human Neuroscience 5* (2011), 41.

[46] NICKERSON, D. History of the munsell color system and its scientific application. *Journal of the Optical Society* (1940).

[47] NITHYANAND, R., SAXENA, N., TSUDIK, G., AND UZUN, E. Groupthink: Usability of secure group association for wireless devices. *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (2010), 331–340. ACM ID: 1864399.

[48] OLLESCH, H., HEINEKEN, E., AND SCHULTE, F. P. Physical or virtual presence of the experimenter: Psychological online-experiments in different settings. *International Journal of Internet Science 1*, 1 (2006), 71–81.

[49] OLMEDO, E. L., AND KIRK, R. E. Maintenance of vigilance by non-task-related stimulation in the monitoring environment. *Perceptual and Motor Skills 44*, 3 (1977), 715–723.

[50] O'MALLEY, J. J., AND POPLAWSKY, A. Noise-induced arousal and breadth of attention. *Perceptual and Motor Skills 33*, 3 (1971), 887–890.

[51] PAUL, C., MORSE, E., ZHANG, A., CHOONG, Y.-Y., AND THEOFANOS, M. A field study of user behavior and perceptions in smartcard authentication. In *Human-Computer Interaction, INTERACT 2011*, vol. 6949 of *LNCS*. Springer Berlin / Heidelberg, 2011, pp. 1–17.

[52] RIVA, G., TERUZZI, T., AND ANOLLI, L. The use of the internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology & Behavior 6*, 1 (2003), 73–80.

[53] ROBLES, R. J., KIM, T.-H., COOK, D., AND DAS, S. A review on security in smart home development. *International Journal of Advanced Science and Technology 15* (2010).

[54] ROGERS, R. D., AND MONSELL, S. Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General 124*, 2 (1995), 207.

[55] SCHNEIER, B. Two-factor authentication: Too little, too late. *Communications of the ACM 48*, 4 (2005), 136.

[56] SÖDERLUND, G., ET AL. Positive effects of noise on cognitive performance: Explaining the moderate brain arousal model. In *The 9th Congress of the International Comisssion on the Biological Effects of Noise* (2008), Leibniz Gemeinschaft, pp. 378–386.

[57] STUDER, A., PASSARO, T., AND BAUER, L. Don't bump, shake on it: The exploitation of a popular accelerometer-based smart phone exchange and its secure replacement. In *Proceedings of the 27th Annual Computer Security Applications Conference* (2011), ACM, pp. 333–342.

[58] TUYLS, P., AKKERMANS, A. H., KEVENAAR, T. A., SCHRIJEN, G.-J., BAZEN, A. M., AND VELDHUIS, R. N. Practical biometric authentication with template protection. In *International Conference on Audio-and Video-Based Biometric Person Authentication* (2005), Springer, pp. 436–446.

[59] UR, B., JUNG, J., AND SCHECHTER, S. The current state of access control for smart devices in homes. In *Workshop on Home Usable Privacy and Security (HUPS)* (2013), HUPS 2014.

[60] UZUN, E., KARVONEN, K., AND ASOKAN, N. Usability analysis of secure pairing methods. In *Financial Cryptography and Data Security*, S. Dietrich and R. Dhamija, Eds., vol. 4886 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2007, pp. 307–324.

[61] VAUDENAY, S. Secure communications over insecure channels based on short authenticated strings. In *Annual International Cryptology Conference* (2005), Springer, pp. 309–326.

[62] VON AHN, L., BLUM, M., HOPPER, N. J., AND LANGFORD, J. CAPTCHA: Using hard AI problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques* (2003), Springer, pp. 294–311.

[63] WEIR, C. S., DOUGLAS, G., CARRUTHERS, M., AND JACK, M. User perceptions of security, convenience and usability for ebanking authentication tokens. *Computers & Security 28*, 1-2 (2009), 47–62.

[64] WHELAN, R. Effective analysis of reaction time data. *The Psychological Record 58*, 3 (2008), 475–482.

[65] WORTMANN, F., AND FLÜCHTER, K. Internet of Things, technology and value added. *Business & Information Systems Engineering 57*, 3 (2015), 221–224.

[66] WYSZECKI, G., AND STILES, W. S. *Color Science*, vol. 8. Wiley New York, 1982.

[67] XIA, F., YANG, L. T., WANG, L., AND VINEL, A. Internet of Things. *International Journal of Communication Systems 25*, 9 (2012), 1101–1102.

[68] ZURKO, M. E. User-centered security: Stepping up to the grand challenge. In *Computer Security Applications Conference, 21st Annual* (2005), IEEE, pp. 14–pp.