User Errors in the Performance of Security Critical Tasks in the Context of Adversarial Noise

Tyler Kaczmarek
University of Californa, Irvine
tkaczmar@uci.edu

Alfred Kobsa
University of Californa, Irvine
kobsa@uci.edu

Robert Sy
University of Californa, Irvine
robertsyproductions@yaho.com

Gene Tsudik
University of Californa, Irvine
gene.tsudik@uci.edu

arXiv:1406.3002v1 [cs.HC] 21 May 2014

ABSTRACT

User errors that occur while performing security-critical tasks can lead to undesirable or even disastrous consequences. The complexity of such tasks is one major factor influencing mistakes and failures. Prior research considered this topic extensively. However, another important issue that has not been explored is the impact of both accidental and intended distractions on users performing security-critical tasks. In particular, it is unclear whether, and to what extent, various distractions (e.g., audio or visual) can influence user behavior and/or trigger mistakes.

In this paper, we report on a pilot study to assess and analyze the impact of unexpected audio distractions. Conducting a thorough and meaningful study requires recruiting a large number of participants. To reduce complexity and potential experimenter-induced bias of conducting such a study, we experimented with a novel fully automated and unattended setup. We employed this setup to administer a range of unexpected (yet naturally occurring) sound effects while study participants attempted to perform a security-critical task. We report on some surprising findings from our experiments and discuss various consequences as well as promising research directions for future work.

1. INTRODUCTION

The physical world is a noisy and distracting place, where truly quiet or sterile environments are quite rare. Most people are accustomed to some degree of audio and visual distraction in their everyday lives. However, human behavior might be influenced in 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 [Goodrich et al. 2009]. In device pairing, one needs to compare things (e.g., numbers, text, pictures, or sounds), or perform some physical task over an "out of band" (OOB) channel [Kobsa et al. 2009].

All these tasks require some form of human involvement, which represents the weakest link and determines overall security [Goodrich et al. 2009, Kobsa et al. 2009, Nithyanand et al. 2010, Kainda et al. 2009, Kainda et al. 2010, Kobsa et al. 2013, Paul et al. 2011]. 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 [Nithyanand et al. 2010, Kainda et al. 2009], conducted by researchers from two groups:

The first is mostly comprised of security experts, who focus on developing methods that provide an acceptable level of security and mitigate the probability of users making security-related mistakes [Kainda et al. 2010]. Unfortunately, since most security experts are not usability or HCI experts, methods developed by this group are secure, yet often burdensome for non-specialist users [Nithyanand et al. 2010, Kainda et al. 2010].

The second group is mostly composed of usability experts. Not surprisingly, their work exhibits the opposite problem: Security-related mistakes are often ignored in favor of examining the usability of protocols under (nearly) ideal conditions. Consequently, these usability studies do not consider the threat of deliberate deception, such as causing a user to validate a transaction that is an order of magnitude more expensive than the true price, due to a deliberate shift of the tiny, hard-to-see decimal point one place to the right. While it is obvious that such deception could have grave security and economic implications, little work has been done

to assess this threat.

One major reason for the dearth of prior work in this area is the difficulty of conducting traditional user experiments. Since human errors are relatively rare in these cases, 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 for 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. Though our original expectation was for such an unexpected audio interference to have a negative impact on the completion of security-critical tasks, the noise actually had a facilitatory effect. We were able to successfully test 125 subjects in this environment, without anyone actively instructing or monitoring them, and without incurring high overhead traditionally associated with running so many trials.

Organization: Section 2 describes related work. Then, Section 3 presents the design and setup of our experiments. Section 4 provides a brief overview of statistical methods used, followed by Section 5 which presents experimental results. Next, Sections 6 and 7 summarize lessons learned from this experience and discuss conclusions, respectively. Section 8 acknowledges certain limitations of our approach. Then, Section 9 addresses ethical considerations and Section 10 concludes the paper with the discussion of future work.

2. RELATED WORK

This section overviews related work in three areas: (1) automated experiments, (2) human-assisted security methods, and (3) effects of noise on human task performance.

2.1 Automated Experiments

To the best of our knowledge, there has been no prior usability study utilizing a fully automated (unattended) experimental setup with a video-projected experimenter. However, there have been precedents with virtually attended experiments and unattended online surveys, in many fields, most notably, psychology. There is a sizable body of work supporting validity and precision of such unattended online experiments, as compared to more traditional attended experiments in a lab setting. In particular, [Ollesch et al. 2006] found no significant difference in psychometric data collected from an attended experiment in a lab setting and its online, virtually attended counterpart. This is further supported by [Riva et al. 2003] in the comparison of data collected from unattended online and attended offline questionnaires. Guidelines for creating the best possible draw from the intended population base are provided by [Birnbaum 2004]. Finally, [Lazem and Gracanin 2010] replicated two classical social psychology experiments, where the experimenter and the three participants, instead of being physically co-present, were represented as avatars in Second Life. The outcomes were found to be very similar.

However, all aforementioned studies assume that an unattended or virtually attended experiment occurs online. There appears to be no prior work involving an offline, unattended experimental setup.

2.2 User Studies of Security Protocols

Both security and usability experts have extensively studied the common setting of secure device pairing. In it, wireless devices have no prior knowledge of one another and, hence, there is no pre-existing security context. This is further complicated by the inability to rely on either a common Public Key Infrastructure (PKI) or a mutually Trusted Third Party (TTP). This accentuates threats of man-in-the-middle (MitM) or "evil twin" attacks during the pairing protocol. Consequently, involvement by a human user has been proposed, in order to verify (over a low-bandwidth OOB channel) message integrity of the protocol that transpires over the normal channel.

For example, Short Authentication String (SAS) protocols require a user to compare two short strings, of about 20 bits each [Laur et al. 2005]. Relative difficulty of fast, accurate completion of this task by human users led to the development of alternative protocols.

The first usability study of pairing techniques was carried out by [Uzun et al. 2007]. They determined that the most accurate way to compare a pair of SAS was in the "compare and confirm" method, wherein the user would be presented a SAS by both of the machines they are trying to pair, and would be to asked to confirm whether or not the two SASs match.

[Goodrich et al. 2009] introduced an authentication technique that utilized a "Mad-Lib" type structure, where the participating devices, based on the protocol outcome, compose a nonsensical phrase out of several short English words. The human user is then tasked with determining whether the two devices came up with matching phrases. This technique was found to be easier to complete by non-specialist users.

[Kobsa et al. 2009] reported on a comprehensive comparative usability study of eleven major secure device pairing methods, measuring task performance times, task completion rates, perceived security and perceived usability. The main outcome was the grouping of the investigated methods into three clusters, following a principal components analysis.

[Kainda et al. 2009] examined usability of device pairing

in a group setting, where up to 6 users tried to connect their devices to one another, and found that group effort decreased the expected rate of security and non-security failures. Although, an inherent "insecurity of conformity" was also identified, wherein users would deliberately lie about an observed string in order to "fit in" with the majority opinion of a group.

[Nithyanand et al. 2010] also examined the pairing of multiple devices in a group setting with groups of 4 or 6 members. They found that groups are nearly immune to insertion attacks, where an adversary will pretend to be a member of the group, and thus change the expected SAS for all members. They also found that groups are particularly vulnerable to a modified man-in-the-middle attack where a single member of the group is given false information, and instead of rejecting their incorrect SAS, they conform to the positive result the rest of the group reports.

[Gallego et al. 2013] found that performance on out-of-band tasks in secure device pairing could be improved through the addition of a score metric on the user's performance, resulting in a considerable reduction in both safe and fatal errors.

There are numerous other results in the area of secure wireless device pairing. However, most of them focused on newly proposed techniques rather than on comparative usability.

2.3 Impact of Noise on Task Performance

In the field of psychology, there are conflicting results with respect to the influence of noise on human task performance. Some experiments claim a positive effect [Hockey 1970, O'Malley and Poplawsky 1971, Baker and Holding 1993], while others report exactly the opposite [Childs and Halcomb 1972, Benignus et al. 1975]. Several explanations of this phenomenon have been proposed. Initially, it was thought that the type of noise used was the primary factor that would cause either an inhibitory or facilitatory effect. [Hockey 1970] demonstrated though that this is not the case, as studies utilizing a diverse range of audio conditions have reported both possible outcomes.

The subsequent explanation was that task complexity might have something to with the effect of noise upon the completion of that task. [Hockey 1970, Benignus et al. 1975] have shown that there may be such a causal relationship. Task complexity is defined by the task's event rate (i.e., how many elements of the task are received within a given period of time), and by the number of different sources from which these task elements are received. Tasks that have a low event rate and a low number of sources are more likely to be facilitated than impaired by the introduction of noise. In contrast, tasks with high event rates and many sources are more likely to be impaired by noise. This hypothesized relationship views noise as a general stimulant that heightens general sensory arousal: if a subject is at a very low level of arousal before the introduction of the noise, it can help sharpen their focus and improve task performance [Olmedo and Kirk 1977, Koelega and Brinkman 1986]. However, for a subject who is already at a high level of sensory arousal, the added stressor can overload them and introduce errors in task completion [Childs and Halcomb 1972, Harris 1960].

Furthermore, [O'Malley and Poplawsky 1971] showed that noise can affect behavioral selectivity. This means that while noise may not have a consistent positive or negative impact on task completion in all cases, noise may consistently have a negative effect on tasks that require the subject to detect signals in their periphery, and noise may have a consistent positive effect on task completion when the subject has to focus on signals coming from the center of their field of attention. This suggests that, regardless of task complexity, the addition of noise may narrow a subject's area of attention.

3. EXPERIMENT

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

3.1 Apparatus

The setting for our study was carefully designed to facilitate fully automated experiments with a variety of sensory input. 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 1(a) shows the experiment location generally, from the side, and Figure 1(b) shows the setup from the subject's perspective (front view), The setup 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), a motion detector, as well as controllable lights and electricity outlets. The Smartboard is an interactive whiteboard (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 3.

This setup allows for a fully unattended experiment. The only (and strictly off-line) involvement of an experimenter amounted to infrequent re-calibration of sound effects and repair of some components that suffered (minor) damage throughout the study.

3.2 Procedures

The goal of the experiment was to measure user errors when attempting to pair two wireless devices via Bluetooth, while being exposed to potentially distracting and possibly "malicious" sound effects. To make the "attack" less noticeable, we use three sounds that can be normally encountered in both open and enclosed public spaces: (1) baby crying, (2) hammer striking a wall, and (3) helicopter rotors spinning.



Figure 1: Experimental Setup: (a) Side view (speakers over the door), (b) Front view

Reasons for selecting three of these specific sounds as audio stimuli are discussed in Section 8.2 below.

The sounds were played at normal volume from a set of speakers behind the subject. Specific volumes of the three sounds (measured at a typical subject's position) were as follows:

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

Even the highest of these three volumes (80 dB) is well within the safe range, as defined by the US Occupational Safety & Health Administration (OSHA) guidelines.[1]

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 nearby office and it is not visible to the subject; refer to Figure 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.

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 four events occurs: either silence is maintained throughout the experiment, or one of the aforementioned three 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 window, until they get it correct. 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 4), thus allowing us to email the promised Amazon coupon as a participation reward.

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 re-trials, task success rates, and task completion times, as well as a video recording of the entire encounter, as shown in Figure 5.

3.3 Hypotheses

Our two initial hypotheses were that, if an unexpected and potentially distracting noise is introduced while an unsuspecting subject attempts to to pair two Bluetooth devices:

H1 We would observe a different (higher) error rate
and
H2 The pairing process would take a different (longer) amount of time to complete successfully,

than in the same setting without any noise interference.

3.4 Subjects

[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. See: `https://www.osha.gov/SLTC/noisehearingconservation/`

In prior studies on usability of pairing protocols with a human in the loop [Goodrich et al. 2009], [Nithyanand et al. 2010], [Kainda et al. 2009], 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 three 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 125 subjects, there were 82 males and 43 females. Most of them (117 out of 125) appeared to be college-aged (18-24 years), while 8 seemed to belong to a somewhat older group (30+ years). This demographic breakdown is reflective of, and influenced by, the experiment's nature and location. Since we solicited participants passively and 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.

4. STATISTICAL TOOLS USED

We briefly overview statistical tools utilized in the analysis of collected data. Given some familiarity with Barnard's exact test, this section can be skipped with no lack of continuity.

4.1 Barnard's Exact Test

[Barnard 1947] specifies a method for testing independence of rows and columns in a contingency table that takes into account nuisance parameters, which are auxiliary parameters that may not be of immediate interest, but instead have a direct impact on the parameter(s) being evaluated. Barnard's exact test seeks to maximize the impact of any such nuisance parameters in maximizing p, the likelihood that the two rows in the contingency table are samples taken from the same population. This means that the p values found by this test are worst-case estimates, and are more suitable for extreme cases than similar methods, and are at least as powerful in non-extreme cases.

4.2 Pairwise examination

In case of multiple comparisons, researchers can assign an acceptable type I error α (false positive) either to each individual comparison or jointly, across all comparisons. As advised by [Hochberg and Tamhane 1987], if "...inferences are unrelated in terms of their content or intended use (although they may be statistically dependent), then they should be treated separately and not jointly." This position has been adopted in this study. To judge the results along a joint α, a significance level $\frac{\alpha}{n}$ can be chosen for each test, with n being the number of pairwise comparisons performed. This corresponds to the so-called Bonferroni correction.

5. RESULTS

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

5.1 Data Cleaning

Subject data was discarded in three cases. First, we removed the instances in which participants arrived either in pairs or larger groups. Their data were eliminated in order to mitigate any effect on subject performance that could be skewed due to social facilitation. It has been shown that being under the observation of others can have a positive impact on subjects performing tasks of low levels of complexity [Aiello and Douthitt 2001]. Second, there was a number of participants with old-style flip phones. Such older phones were technically unable to establish a Bluetooth connection with our client.

Our experiment 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 is one without substantial hearing or visual impairment. However, due to the unattended nature of our experiment, we could not proactively rule out any such candidate 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 the subjects that fit our criteria, which could adversely influence the accuracy of collected data. Therefore, during later review of subjects' video-recorded experiments, we made an attempt to rule out any subject with obvious visual and/or hearing impairments.

Finally, in discussing our results below, we differentiate between pairings and failures that occurred in the first two-minute trial window, and those that occurred across all attempts. While the differences are not large, the former results may capture those subjects better that have already had some pairing experience in the past.

5.2 Task Completion Rate

Table 1 shows the numbers of subjects whose first attempt at pairing resulted in a success and failure, respectively, plus the failure rate for each stimulus (as well as control) condition.

The immediate observation is that these results clearly refute our hypothesis [H1] in Section 3.3 as the failure rate markedly decreases with the introduction of noise. Section 7 discusses this further.

Table 2 displays the pairwise differences in subject failure rates between the control and the three stimulus conditions. Barnard's exact test shows that the differences are statistically significant ($p < 0.05$) with respect to all three stimuli.

Table 3 shows the total number of pairing attempts that ended in success (as well as those that failed) across *all*

Table 1: Subject failure rate, first attempt only

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
Total	107	18	0.14

Table 2: Barnard's Exact Test on subject failure rates between control and each stimulus

Stimulus	Total Pairings	Failure Rate	Nuisance Parameter	p
None(control)	40	0.34	–	–
Baby	24	0.04	0.95	0.03
Hammering	36	0.08	0.91	0.01
Helicopter	25	0.04	0.89	0.01

pairing trials, and the failure rate for each stimulus (and control) condition. We note that, not every subject who initially failed chose to re-try. However, every subject who re-tried was successful the second time.

Table 3: Failure rate by stimulus across all attempts

Stimulus	#Successful Pairings	#Failed Pairings	Failure Rate
None (control)	28	13	0.39
Baby	24	1	0.04
Hammering	34	3	0.11
Helicopter	24	1	0.04
Total	110	18	0.16

We also partitioned subject failure rates by gender. While Table 4 seems to indicate that female subjects were substantially less likely to fail on the initial attempt than their male counterparts, preforming Barnard's exact test on the differences in subject failure rate between men and women revealed that $p = 0.14$, which is not statistically significant.

5.3 Task Completion Times

Table 4: Subject failure rate by gender, first attempt only

Gender	#Successful Subjects	#Unsuccessful Subjects	Failure Rate
Male	67	15	0.18
Female	39	4	0.09

Table 5 shows average completion times in successful trials for subjects under each stimulus. As can be seen from these results, there does not appear to be a statistically significant difference in completion time between the control condition and each of the noise conditions. However, the difference between the hammering condition and the control condition approaches statistical significance ($p = 0.07$).

Table 5: Avg times (sec) for successful pairing

Stimulus	Mean Time	Standard Deviation	t-value wrt control	p
None	34.41	13.78	–	–
Baby	31.13	10.06	0.97	0.35
Hammering	28.82	9.76	1.84	0.07
Helicopter	31.33	13.13	0.81	0.39

As with subject failure rates, we also examined the subjects' completion times for successful pairing attempts by gender. Though there may appear to be a noticeable difference in the average completion times for male and female subjects, as illustrated by Table 6, a pairwise t-test shows that the difference in completion times with respect to the subjects' gender is statistically insignificant ($p = 0.29$).

Table 6: Avg times (sec) for successful pairing by gender

Gender	Mean Time	Standard Deviation
Male	30.57	10.92
Female	33.21	13.85

6. 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 5.1. 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.

Our experiments were conducted during the academic year while the term was in session. During that time, even though the area where recruitment posters were placed experienced heavy foot traffic, our prominently placed signs did not attract as many subjects as we expected in the short term (1-2 weeks). Therefore, in order to gain a sufficient number of subjects, the experiment lasted about 6 weeks with some short breaks when recruitment posters were removed. These short breaks actually proved useful, since we believe that periodic appearance and re-appearance of recruitment posters motivated additional participants.

In retrospect, video surveillance of the experimental setup proved invaluable, both for actual security purposes and for being able to later correct experimental lapses.

7. DISCUSSION OF OBSERVED EFFECTS

The introduction of several types of peripheral audio noise did not appear to interfere with the completion of the task of Bluetooth pairing. In fact, collected data shows a significant decrease in failure rates for every stimulus. In both the cases of the sounds of a baby crying and a helicopter's rotors spinning, there was only a single failure across 25 attempts, as opposed to the control case, where every third attempt resulted in failure, as shown in Table 1.

This result, while initially unexpected, is actually consistent with the Brain Arousal Model of [Söderlund 2008] as well as the results of [Benignus et al. 1975, Hockey 1970, O'Malley and Poplawsky 1971]. Our experiment has a single, centrally-located source and a low event rate, so it would be reasonable to expect noise to have a positive effect on successful task completion rates [Benignus et al. 1975, Hockey 1970]. Audio signals in our study came from the center of participants' area of attention and not the periphery, i.e., the Smartboard and participants' smartphones were in front of them. This suggests that noise caused participants to narrow their focus of attention [O'Malley and Poplawsky 1971], which might be conducive to better task performance. Nevertheless, our study is novel in the context of security-critical tasks, since, by its very nature, the process of Bluetooth device pairing is significantly shorter in duration than the attention-intensive vigilance tasks discussed in related literature.

Average (successful) task completion times did not change in a statistically significant manner when stimuli were introduced. However, as Table 5 shows, hammering does *approach* statistical significance. One intuitive conjecture for this modestly improved completion time is that it could be caused by the subjects' increased level of sensory arousal due to hammering. It might have had a different effect on completion time than the other stimuli due to the nature of the sound itself. The baby and helicopter stimuli were both continuous sounds that, once started, did not vary much throughout their duration. Whereas, hammering sound was unique in that it was composed of short, discrete, staccato bursts of sound. This more abrupt nature could have facilitated faster task completion.

8. LIMITATIONS

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

8.1 Subjects

We admittedly experimented with a narrow subject group dominated by young and tech-savvy college students. This is a consequence of the specific campus location of the unattended setup. Replicating it in a commercial setting (e.g., an office building) would be possible and useful. However, a really diverse group of participants can only be passively recruited 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, for two reasons: First, our setup involved specialized and expensive equipment security of which 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.

On a related note, 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 portion of the campus of a large university, 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 [Brant and Fozard 1990]. It is quite conceivable that older and/or technologically non-adept people[2] would react differently to our noise stimuli.

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

8.2 Diversity of Stimuli

We experimented with three stimuli through a subjective process of elimination, with the intention of getting as many diverse noise types as we could rigorously 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 125 subjects, the rate of participation decreased markedly over time and ceased completely after 6 weeks. As it turned out, 125 was just enough for three 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 three stimuli to be as diverse as possible: (1) baby crying was a continuous, irregular, human-generated sound, (2) helicopter rotors was a continuous, regular, mechanical sound, and (3) hammering was a discrete, regular mechanical sound.

8.3 Insufficiently Security-Critical Task

We suspect that most participants were unaware, ahead of time, of the purpose and details of our experiment. However, they clearly understood, during the experiment, 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.

8.4 Synthetic Environment

Our unattended experiment setup is clearly very synthetic, for several reasons: First, it is normally very quiet, unlike 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 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.

[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.

8.5 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 Internet provider log-in process.

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.

9. ETHICAL CONSIDERATIONS

Experiments described in this paper 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 but can not be provided here in order to avoid violating anonymous submission rules.

We also 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 3.2.

10. FUTURE WORK

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.

An interesting next step would be to conduct a similar experiment with various visual stimuli, that is, to model an adversary who controls some aspect of the visual environment of the experiment. The natural follow on then would be to investigate the effects of mixed audio-visual stimuli.

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.

Acknowledgments

This research was supported by the Anonymous Sources under grant XXXX. We would like to thank all anonymous participants in our study.

11. REFERENCES

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

[Baker and Holding 1993] Baker, M. A. and Holding, D. H. 1993. The effects of noise and speech on cognitive task performance. *The Journal of general psychology 120,* 3, 339–355.

[Barnard 1947] Barnard, G. A. 1947. Significance tests for 2x2 tables. *Biometrika 34,* 1-2, 123–138.

[Benignus et al. 1975] Benignus, V. A., Otto, D. A., and Knelson, J. H. 1975. Effect of low-frequency random noises on performance of a numeric monitoring task. *Perceptual and motor skills 40,* 1, 231–239.

[Birnbaum 2004] Birnbaum, M. H. 2004. Human research and data collection via the internet. *Annual Review of Psychology 55,* 1, 803–832. PMID: 14744235.

[Brant and Fozard 1990] Brant, L. J. and Fozard, J. L. 1990. Age changes in puretone hearing thresholds in a longitudinal study of normal human aging. *The Journal of the Acoustical Society of America 88,* 2.

[Childs and Halcomb 1972] Childs, J. M. and Halcomb, C. G. 1972. Effects of noise and response complexity upon vigilance performance. *Perceptual and motor skills 35,* 3, 735–741.

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

[Goodrich et al. 2009] Goodrich, M. T., Sirivianos, M., Solis, J., Soriente, C., Tsudik, G., and Uzun, E. 2009. Using audio in secure device pairing. *International Journal of Security and Networks 4,* 1, 57–68.

[Harris 1960] Harris, W. 1960. Ph.D. thesis. Ph.D. thesis, University of Southern California.

[Hochberg and Tamhane 1987] Hochberg, Y. and Tamhane, A. 1987. *Multiple comparison procedures.* Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

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

[Kainda et al. 2009] Kainda, R., Flechais, I., and Roscoe, A. W. 2009. Usability and security of out-of-band channels in secure device pairing protocols. *Proceedings of the 5th Symposium on Usable Privacy and Security*, 11:1–11:12. ACM ID: 1572547.

[Kainda et al. 2010] Kainda, R., Flechais, I., and Roscoe, A. W. 2010. 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*. SOUPS '10. 5:1–5:13. ACM ID: 1837117.

[Kobsa et al. 2013] Kobsa, A., Nithyanand, R., Tsudik, G., and Uzun, E. 2013. Can jannie verify? usability of display-equipped RFID tags for security purposes. *Journal of Computer Security 21,* 3 (Jan.), 347–370.

[Kobsa et al. 2009] Kobsa, A., Sonawalla, R., Tsudik, G., Uzun, E., and Wang, Y. 2009. Serial hook-ups: a comparative usability study of secure device pairing methods. *Proceedings of the 5th Symposium on Usable Privacy and Security*, 10:1–10:12. ACM ID: 1572546.

[Koelega and Brinkman 1986] Koelega, H. S. and Brinkman, J.-A. 1986. Noise and vigilance: An evaluative review. *Human Factors: The Journal of the Human Factors and Ergonomics Society 28,* 4, 465–481.

[Laur et al. 2005] Laur, S., Asokan, N., and Nyberg, K. 2005. Efficient mutual data authentication using manually authenticated strings. Cryptology ePrint Archive, Report 2005/424. `http://eprint.iacr.org/`.

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

[Nithyanand et al. 2010] Nithyanand, R., Saxena, N., Tsudik, G., and Uzun, E. 2010. Groupthink: usability of secure group association for wireless devices. *Proceedings of the 12th ACM international conference on Ubiquitous computing*, 331–340. ACM ID: 1864399.

[Ollesch et al. 2006] Ollesch, H., Heineken, E., and Schulte, F. P. 2006. Physical or virtual presence of the experimenter: Psychological online-experiments in different settings. *International Journal of Internet Science 1,* 1, 71–81.

[Olmedo and Kirk 1977] Olmedo, E. L. and Kirk, R. E. 1977. Maintenance of vigilance by non-task-related stimulation in the monitoring environment. *Perceptual and motor skills 44,* 3, 715–723.

[O'Malley and Poplawsky 1971] O'Malley, J. J. and Poplawsky, A. 1971. Noise-induced arousal and breadth of attention. *Perceptual and motor skills 33,* 3, 887–890.

[Paul et al. 2011] Paul, C., Morse, E., Zhang, A., Choong, Y.-Y., and Theofanos, M. 2011. A field study of user behavior and perceptions in smartcard authentication. In *Human-Computer Interaction âĂŞ INTERACT 2011*, P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, and M. Winckler, Eds. Lecture Notes in Computer Science, vol. 6949. Springer Berlin / Heidelberg, 1–17.

[Riva et al. 2003] Riva, G., Teruzzi, T., and Anolli, L. 2003. The use of the internet in psychological research: comparison of online and offline questionnaires. *CyberPsychology & Behavior 6,* 1, 73–80.

[Söderlund 2008] Söderlund, G. 2008. Positive effects of noise on cognitive performance: Explaining the moderate brain arousal model. In *Noise as a Public Health Problem: Proceedings*. Leibniz Gemeinschaft, 378–386.

[Uzun et al. 2007] Uzun, E., Karvonen, K., and Asokan, N. 2007. Usability analysis of secure pairing methods. In *Financial Cryptography and Data Security*, S. Dietrich and R. Dhamija, Eds. Lecture Notes in Computer Science, vol. 4886. Springer Berlin Heidelberg, 307–324.

APPENDIX

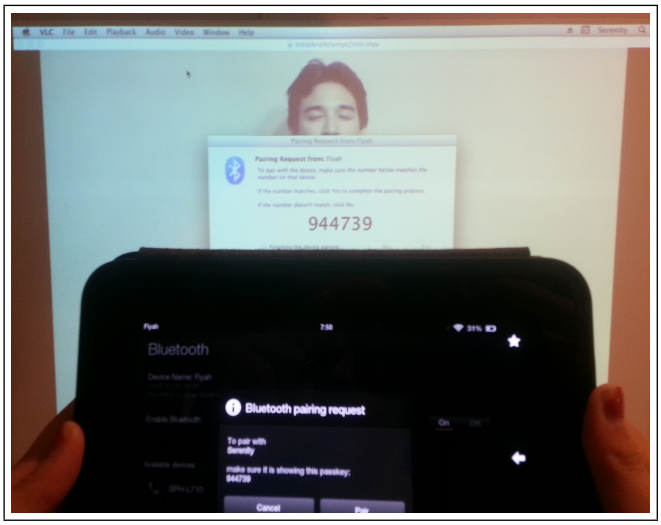


Figure 2: Bluetooth confirmation screen, from subject's perspective



Figure 3: Experimenter proxy giving video instructions

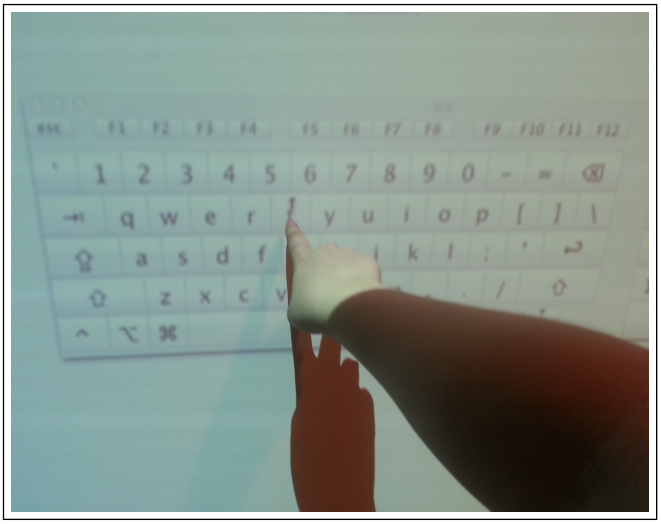


Figure 4: Subject entering email address on Smartboard

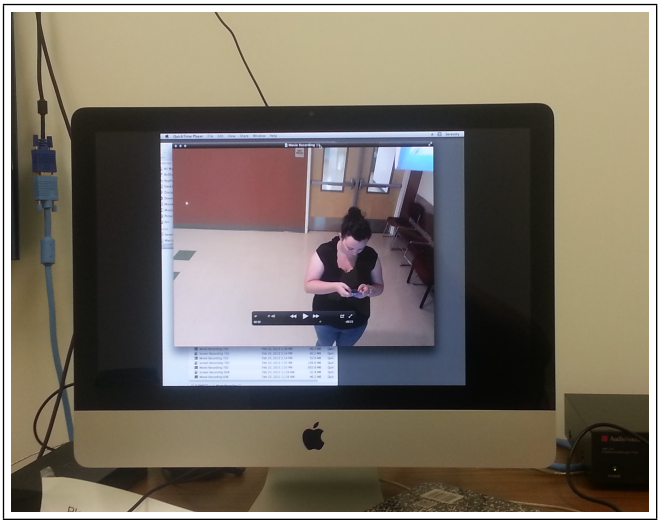


Figure 5: The reviewer's perspective (separate office)