An Unattended Study of Users Performing Security Critical Tasks Under Adversarial Noise

Tyler Kaczmarek
UC Irvine
tkaczmar@uci.edu

Alfred Kobsa
UC Irvine
kobsa@uci.edu

Robert Sy*
DHS
robertsyproductions@yahoo.com

Gene Tsudik
UC Irvine
gene.tsudik@uci.edu

***Abstract*—User errors while performing security-critical tasks can lead to undesirable or even disastrous consequences. One major factor influencing mistakes and failures is complexity of such tasks, which has been studied extensively in prior research. Another important issue which hardly received any attention 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, unexpected sensory cues (e.g., auditory or visual) can influence user behavior and/or trigger mistakes. Better understanding of the effects of intended distractions will help clarify their role in adversarial models. As part of the research effort described in this paper, we administered a range of naturally occurring – yet unexpected – sounds while study participants attempted to perform a security-critical task. We found that, although these auditory cues lowered participants' failure rates, they had no discernable effect on their task completion times. To this end, we overview some relevant literature that explains these somewhat counter-intuitive findings.**

Conducting a thorough and meaningful study on user errors requires a large number of participants, since errors are typically infrequent and should not be instigated more than once per subject. To reduce the effort of running numerous subjects, we developed a novel experimental setup that was fully automated and unattended. We discuss our experience with this setup and highlight the pros and cons of generalizing its usage.

I. 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 [1]. 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 [2].

All these tasks require some form of human involvement, which represents the weakest link and determines overall security [1], [2], [3], [4], [5], [6], [7]. 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 [3], [4], 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.[1] 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 147 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 II describes related work. Then, Section III presents the design and setup of our experiments. It is

*Work done while at UC Irvine.

[1] See Appendix.

USEC '15, 8 February 2015, San Diego, CA, USA

http://dx.doi.org/10.14722/usec.2015.23014

followed by Section IV which presents experimental results. Next, Sections V and VI summarize lessons learned from this experience and discuss conclusions, respectively. Section VII acknowledges certain limitations of our approach. Then, Section VIII addresses ethical considerations and Section IX concludes the paper with the discussion of future work. Appendix A provides a brief overview of statistical methods used and Appendix B demonstrates the unattended experiment setup.

II. 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.

A. 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. [8] 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. [9] 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 [10]. Finally, Lazem and Gracanin [11] replicated two classical social psychology experiments, where the experimenter and the three participants were represented as avatars in Second Life instead of being physically co-present. The outcomes were 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.

B. User Studies of Security Protocols

Both security and usability experts have extensively studied secure device pairing. In this setting, 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 of 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 [12]. Since accurate task completion was found to be relatively difficult for human users, alternative protocols were developed.

The first usability study of pairing techniques was carried out by Uzun et al. [13]. That study determined that the most accurate way to compare a pair of SAS was the "compare and confirm" method, wherein the user would be presented an 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. [1] introduced an authentication technique that utilized a "Mad-Lib" type structure, where 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. [2] 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. [4] 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. [3] 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. [14] 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.

C. 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 [15], [16], [17], while others report exactly the opposite [18], [19]. 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 [15] 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 [15] and Benignus et al. [19] have shown that such a causal relationship may exist. 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 [20], [21]. 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 [18], [22].

Furthermore, O'Malley and Poplawsky [16] 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.

III. Experiment

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

A. 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 1(a) shows the experimental location from the side, and Figure 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), 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.

B. 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 four sounds that can be encountered in real life, both in open and enclosed public spaces: (1) a baby crying, (2) a hammer striking a wall, (3) helicopter rotors spinning, and (4) a circular saw cutting wood. Reasons for selecting these four specific sounds as audio stimuli are discussed in Section VII-B below.

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

- Baby: 67 dB
- Helicopter: 79 dB
- Hammer: 80 dB
- Saw: 78 dB

Even the highest of these four volumes (80 dB) is well within the *safe range*, as defined by the US Occupational Safety & Health Administration (OSHA) guidelines.[2]

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 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 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 window. 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.

Each subject encountered only one condition. Presenting subjects with two conditions would have biased their per-

[2]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/

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

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

C. 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.

D. Subjects

In prior studies on usability of pairing protocols with a human in the loop [1], [3], [4], it was discovered that a subject population of 20-25 per condition being tested was an acceptable size for obtaining statistically significant[3] findings. Since our planned experiment has one condition for each of the four 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.

[3]See Appendix for the definition of statistical significance.

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 147 subjects, there were 102 males and 45 females. Most of them (139 out of 147) appeared to be college-aged (18-24 years), while 8 seemed to belong to a somewhat older group (30+ 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.

IV. Results

We now discuss the results of the study, starting with data cleaning and proceeding to task completion results. Statistical tools used in data analysis are described in the Appendix.

A. 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 [23]. 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.

Finally, in discussing 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.

B. Task Completion Rate

Table I 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 I. 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
Saw	20	2	0.09
Total	127	20	0.14

Table II shows the parameters for the Barnard's exact test applied pairwise 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 (see the Appendix), which leads us to reject hypothesis H1 in Section III-C, since the failure rate significantly decreases with the introduction of noise. Section VI discusses this further.

TABLE II. 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

Table III 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% 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 III. 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

Table IV shows the total number of pairing attempts that ended in success (as well as those that failed) across *all 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 IV. FAILURE RATE BY STIMULUS ACROSS ALL ATTEMPTS

Stimulus	#Successful Pairings	#Failed Pairings	Failure Rate
None (control)	28	13	0.32
Baby	24	1	0.04
Hammering	34	3	0.08
Helicopter	24	1	0.04
Saw	20	2	0.09
Total	130	20	0.13

We also partitioned subject failure rates by gender. While Table V seems to indicate that female subjects were substantially less likely to fail on the initial attempt than their 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 = 2.32, nuisance parameter = 0.98, $p = 0.14$.

TABLE V. SUBJECT FAILURE RATE BY GENDER, FIRST ATTEMPT ONLY

Gender	#Successful Subjects	#Unsuccessful Subjects	Failure Rate
Male	86	16	0.16
Female	41	4	0.09

C. Task Completion Times

Table VI 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

TABLE VI. 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

in completion times between the control condition and each stimulus.

Table VII 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 effect sizes of the stimuli are not statistically significant from 0 since each of the confidence intervals contains 0.

TABLE VII. 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

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 ($t(148) = 1.23$, $p = 0.22$).

TABLE VIII. AVG TIMES (SEC) FOR SUCCESSFUL PAIRING BY GENDER

Gender	Mean Time	Standard Deviation
Male	30.63	10.92
Female	33.23	13.85

V. 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 IV-A. 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 in a short period of time (1-2 weeks) as we had expected. 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.

VI. 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, with no statistically significant difference in the failure rate between different noise stimuli. In cases of baby crying and 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 I.

This result, while initially unexpected, is actually consistent with the Brain Arousal Model of [24] as well as the results of [19], [15], [16]. 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 [19], [15]. 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 [16], 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.

VII. Limitations

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

A. 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 [25]. It is quite conceivable that older and/or technologically non-adept people[4] 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.

B. 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 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.[5]

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

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

C. 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.

D. 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.

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

[5] 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.

E. 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.

VIII. 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. 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 III-B.

IX. 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 NSF grant CNS-0831526. We would like to thank all anonymous participants in our study.

References

[1] M. T. Goodrich, M. Sirivianos, J. Solis, C. Soriente, G. Tsudik, and E. Uzun, "Using audio in secure device pairing," *International Journal of Security and Networks*, vol. 4, no. 1, pp. 57–68, 2009.

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

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

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

[5] R. Kainda, I. Flechais, and A. W. Roscoe, "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, pp. 5:1–5:13, 2010. ACM ID: 1837117.

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

[7] C. Paul, E. Morse, A. Zhang, Y.-Y. Choong, and M. Theofanos, "A field study of user behavior and perceptions in smartcard authentication," in *Human-Computer Interaction, INTERACT 2011*, vol. 6949 of *LNCS*, pp. 1–17, Springer Berlin / Heidelberg, 2011.

[8] H. Ollesch, E. Heineken, and F. P. Schulte, "Physical or virtual presence of the experimenter: Psychological online-experiments in different settings," *International Journal of Internet Science*, vol. 1, no. 1, pp. 71–81, 2006.

[9] G. Riva, T. Teruzzi, and L. Anolli, "The use of the internet in psychological research: comparison of online and offline questionnaires," *CyberPsychology & Behavior*, vol. 6, no. 1, pp. 73–80, 2003.

[10] M. H. Birnbaum, "Human research and data collection via the internet," *Annual Review of Psychology*, vol. 55, no. 1, pp. 803–832, 2004. PMID: 14744235.

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

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

[13] E. Uzun, K. Karvonen, and N. Asokan, "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*, pp. 307–324, Springer Berlin Heidelberg, 2007.

[14] A. Gallego, N. Saxena, and J. Voris, "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*, pp. 60–68, Springer Berlin Heidelberg, 2013.

[15] G. R. J. Hockey, "Effect of loud noise on attentional selectivity," *The Quarterly Journal of Experimental Psychology*, vol. 22, no. 1, pp. 28–36, 1970.

[16] J. J. O'Malley and A. Poplawsky, "Noise-induced arousal and breadth of attention," *Perceptual and motor skills*, vol. 33, no. 3, pp. 887–890, 1971.

[17] M. A. Baker and D. H. Holding, "The effects of noise and speech on cognitive task performance," *The Journal of general psychology*, vol. 120, no. 3, pp. 339–355, 1993.

[18] J. M. Childs and C. G. Halcomb, "Effects of noise and response complexity upon vigilance performance," *Perceptual and motor skills*, vol. 35, no. 3, pp. 735–741, 1972.

[19] V. A. Benignus, D. A. Otto, and J. H. Knelson, "Effect of low-frequency random noises on performance of a numeric monitoring task," *Perceptual and motor skills*, vol. 40, no. 1, pp. 231–239, 1975.

[20] E. L. Olmedo and R. E. Kirk, "Maintenance of vigilance by non-task-related stimulation in the monitoring environment," *Perceptual and motor skills*, vol. 44, no. 3, pp. 715–723, 1977.

[21] H. S. Koelega and J.-A. Brinkman, "Noise and vigilance: An evaluative review," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 28, no. 4, pp. 465–481, 1986.

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

[23] J. R. Aiello and E. A. Douthitt, "Social facilitation from triplett to electronic performance monitoring.," *Group Dynamics: Theory, Research, and Practice*, vol. 5, no. 3, pp. 163–180, 2001.

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

[25] L. J. Brant and J. L. Fozard, "Age changes in puretone hearing thresholds in a longitudinal study of normal human aging," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 813–820, 1990.

[26] G. A. Barnard, "Significance tests for 2x2 tables," *Biometrika*, vol. 34, no. 1-2, pp. 123–138, 1947.

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

APPENDIX

APPENDIX A: STATISTICAL TOOLS

This section overviews simple statistical tools utilized in the analysis of collected data.

A. Statistical significance

A statement of statistical significance is a statement of confidence that two sets of observations represent samples taken from different populations. This claim is constructed by comparing the probability that two sets of observations could have been taken from the same population distribution. The probability p of the two samplings being from the same population is then evaluated against a low-threshold, α. If $p \leq \alpha$, the result is said to be *statistically significant*, and one could confidently claim that the two samples represent two different populations. We fix the low-threshold at $\alpha = 0.05$.

B. Barnard's Exact Test

Barnard [26] 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. We utilize Barnard's exact test in the examination of subject failure rates.

C. Paired t-tests

The t-test is a method for determining if the differences between two data sets are statistically significant through the examination of their mean and standard deviation. For populations that can be assumed to follow a normal distribution, these two parameters are sufficient for constructing p. A t-test is said to be paired if the two data sets consist of pairs of matched units. We utilize pairwise t-tests in the examination of subject completion times.

D. 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. 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 [27].

APPENDIX B: UNATTENDED EXPERIMENT SETUP

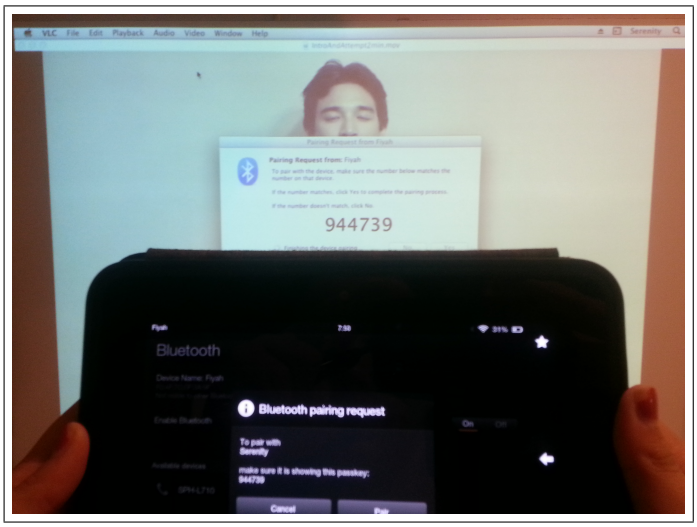

Fig. 2. Bluetooth confirmation screen, from subject's perspective

Fig. 3. Experimenter proxy giving video instructions

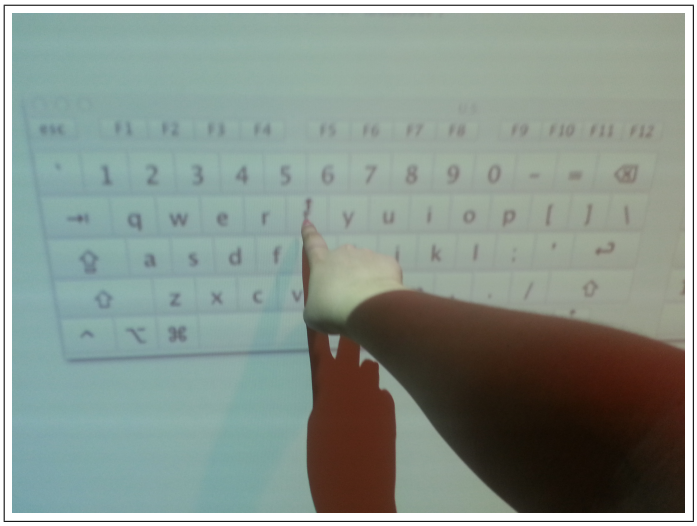

Fig. 4. Subject entering email address on Smartboard

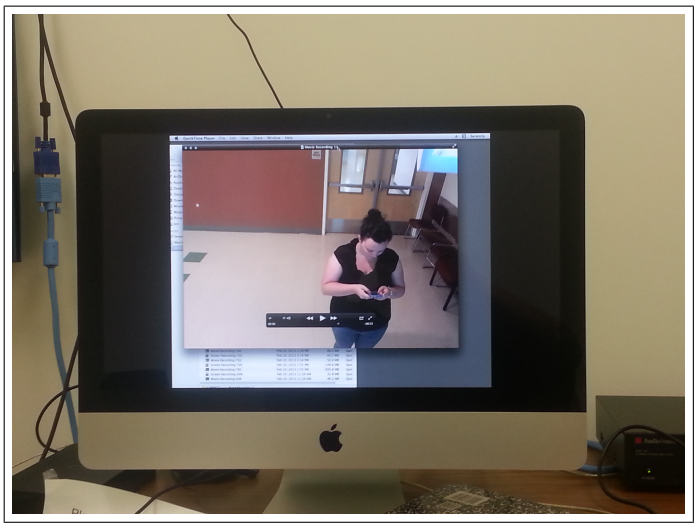

Fig. 5. Post-experimental review of video recordings (separate office)