A Short Study on Security Indicator Interfaces

Technical Report UCSC-WASP-15-04
November 2015

D J Capelis
mail@capelis.dj

Working-group on Applied Security and Privacy
Storage Systems Research Center
University of California, Santa Cruz
Santa Cruz, CA 95064
http://wasp.soe.ucsc.edu/

This work was supported by

Abstract

This technical report documents findings from a preliminary study on how users react to security indicators from a trusted hardware output channel. We recorded how a mixture of test messages affected users' interactions with a simulated web form asking for account credentials. In this report we present the data and draw some conclusions on how this experiment informs trusted platform interface design, and specifically report findings relevant to the interface from our *Lockbox* project.

1 Introduction

End-users frequently disregard, respond erratically or don't even notice the interfaces that are supposed to help them protect their machine. Here we examine hardware interfaces that provide end-users with an auxiliary security screen which displays information from a hardware security system. We are interested in finding out how users react to such a screen and determine whether or not such an interface allows user to make good security decisions.

I designed an experiment where users were prompted to log in to a series of login forms. The screen displayed various iterations of text and we were able to test user's error rates as well as response times to various security displays. I observed both the timing information and the error rates of each screen. With statistical analysis, we draw several conclusions from the data. These supported some of our interface design assumptions, provide insight into phishing type attacks on these types of interfaces and suggest how the hardware should respond when there is no trusted channel present. These conclusions validated portions of the *LockBox* design as well as suggest possible tweaks for further accuracy in the future.

2 Experimental Design

There are two very important metrics that determine success for a security system like this:

Accuracy The user's ability to come to the correct conclusion.

Cognitive Overhead How much thinking the user is required to do to come to a conclusion.

While many, might consider the first metric to be most important, the second can be just as, or more, important. The amount of cognitive overhead a user experiences *determines their willingness to use the system at all.* Since the accuracy rate of a system that a user doesn't use is zero, it could be necessary to trade accuracy for lower cognitive overhead. That said, this data does not support the existence of a tradeoff between these two variables in this particular interface. Our results are consistent with the hypothesis that interfaces with lower cognitive overhead are also more accurate.

Accuracy metrics were gathered by recording whether a user reached the correct decision. Tracking cognitive overhead was more difficult. Since there no neuroscientist or MRI machine was present in our lab, I recorded time between responses. This seems like a reasonable metric for cognitive overhead, but simply observing how quickly someone reacts is an indirect measure.

2.1 Subjects

20 subjects participated in the study. They were recruited from the University of California, Santa Cruz campus and Silicon Valley. The academic participants from UCSC's Computer Science department consisted of one professor, one undergraduate and five graduate students. In addition, two undergraduates and four graduate students participated from other departments at UCSC. Another was studying an unrelated field at Stanford. The remaining six participants worked full time for various Silicon Valley tech companies.

The subjects ranged from 20 to 50 years of age. All were familiar with technology, but not all had chosen it as their vocation. Most were students. In general, the subjects were both young and very familiar with computers.

2.2 Pilot Study

The first six participants were comprised the pilot subjects for the project. When it was found the pilot study was yielding valid data and the methodology appeared to be sufficient, we incorporated data from these subjects into our main study. The methodology did not change after the pilot study. At the conclusion, we verified that the pilot study participants did not produce significantly different results than the remaining users.

2.3 Test Procedure

The test began by providing the subject with a series of instructions. Included in these instructions were information on the test, the methodology and the format. Users were given a username and password and shown an exact picture of contents of the screen they should accept. The users then moved on to a series of login forms where they were either able to login, or press a button that stated the form was insecure. The auxiliary security screen would change at each page and the user's responses where tracked. About 2/3rds of the way through the test, users were told that instead of different types of screens, they would only see two screens: the proper screen and a screen that said the input was insecure. At the conclusion of the test, the users provided feedback on a paper survey.

In the first part of the study, the auxiliary screen displayed one of the following readouts during the study:

- A screen that read `firefox`. Users were instructed that this screen meant their input was secure.
- A screen that read `Firefox`. With the first letter capitalized. Users were instructed to press the button that said the form was insecure if the auxiliary security screen deviated from the lower case "firefox" readout.
- A screen that read `f1refox`. With the first 'i' replaced by a 1.
- A screen that read `internet explorer`. The name of a complete different, but related program.
- A blank screen.

In the second half of the test, users were given a screen that read out `-INPUT INSECURE-`. This interface of affirming an insecure state was compared to previous results where the insecure state was discovered by the lack of a proper security notification on the auxiliary screen.

During the experiment, the subject sat at a computer terminal with a researcher seated behind them jotting observations on a post-it. The researcher had a view of the screen, the auxiliary screen and the user's input. Users did not appear to pick up on any signals from the observer as several thought they were doing badly when they were doing well or well when they were doing badly. The observer helped facilitate this neutrality by wearing a somber expression and a white lab coat.

3 Data

The first 10 questions were screened out of the data as a training period. Since the user responded slowly at first and slowly grew better through the first 10 responses, these responses were screened out as not representative. In addition, we also screened out outliers where users took greater than 30 seconds. This happened only on a few datapoints. The observation notes confirmed the majority of these instances occurred when the subject paused during the test to ask a question or receive clarification.

3.1 Cognitive Overhead Metrics

Cognitive overhead was based on a measure of time between responses. We present both the average responses in graph form as well as a few statistically significant statements that can be made about users cognitive overheads in one part of the experiment vs. another.

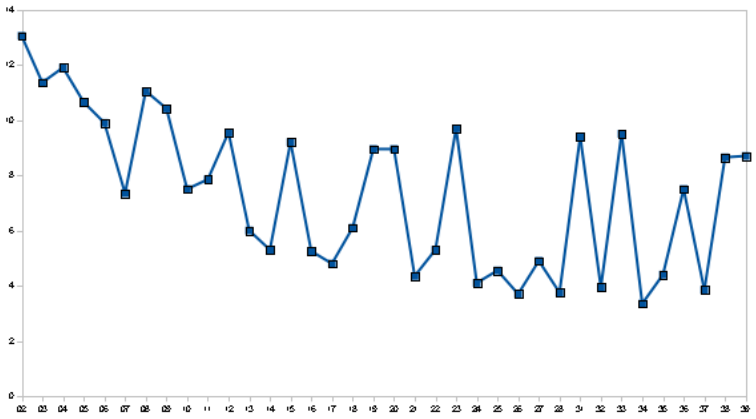

Figure 1: Graph of the averages

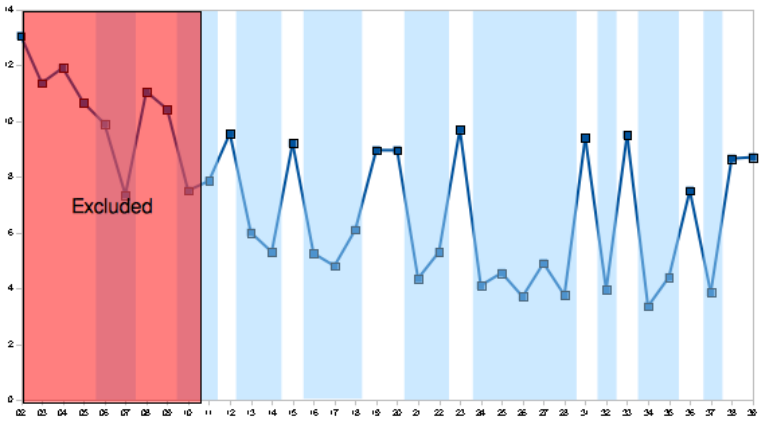

Figure 2: Graph of the averages, showing negative user responses shaded in blue and the excluded section of the results

The following differences were all statistically significant:

- Users were quicker to respond during the phase of the test when the insecure state was affirmatively presented. (p=.029)
- Users were quicker to determine that the system was insecure when presented with an explicit warning instead of a blank screen. (p=.017)
- Users were quicker to respond after the training period. (p¡.01)

3.2 Accuracy Rates

We present both raw and adjusted accuracy rates. The adjusted rates exclude data from three participants who produced particularly erratic and noisy data. These three participants either

didn't correctly understand the experiment or mistakenly assumed that one of the fake screens was correct and the right screen was fake through all or a portion of the experiment. (People seemed inclined to prefer a capital F in their Firefox and would actually begin reject the lowercase one even though they were prompted to use that instead.) These types of mistakes would be unlikely to occur in a design like *LockBox*'s, so we excluded these particular cases in the adjusted results.

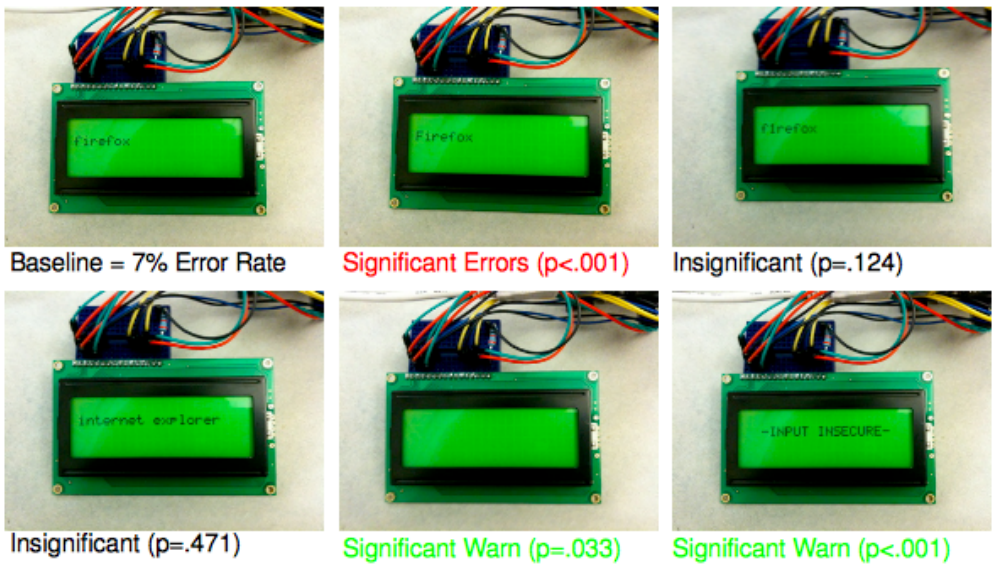

Figure 3: The accuracy rate compared to baseline for each security screen using raw data

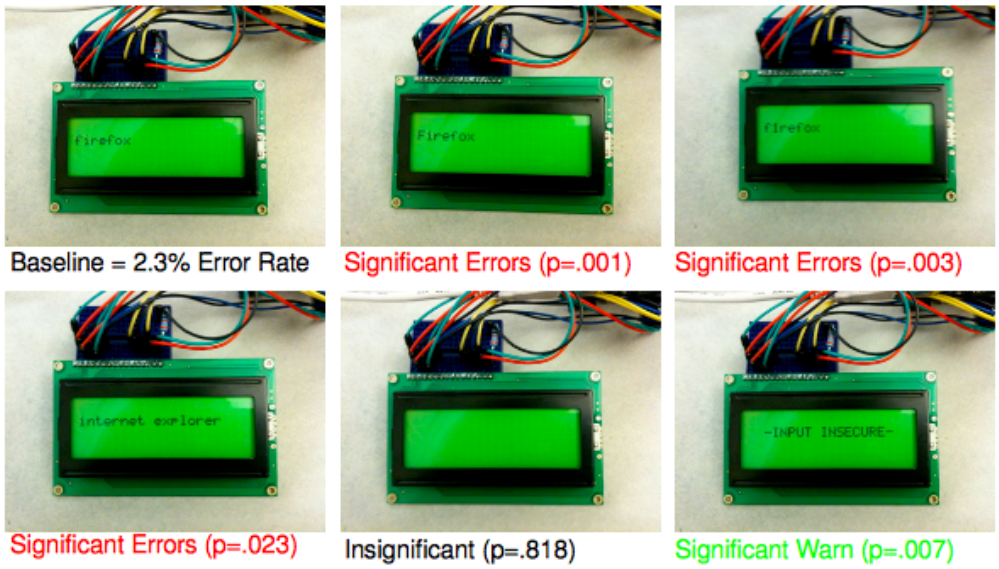

Figure 4: The accuracy rate compared to baseline for each security screen using adjusted data

4 Findings

While security interfaces are tricky and often involve high rates of user error, our findings yield two different findings security interfaces should take into account to increase their success rate.

First, avoid interfaces which allow attackers to pick any strings on the security indicator. Both the accuracy and cognitive overhead data that users were faster and more accurate when they did not have to distinguish between various similar looking screens to determine which one was correct. This data supports the design approach made in *LockBox* to present users only with strings that they chose if a trusted channel is secure, or none at all. This design decision sets the interface in *LockBox* apart from interfaces which allow attackers to present users with similar name, such as in browser address bars in web browsers. The problems and challenges that present themselves for that type of interface are much greater.

Second, when the system is in an insecure mode, affirmatively display this information to the user. A blank screen was less effective than an explicit message. Users were both faster and more accurate if the screen displayed an explicit warning that their input was insecure. This is not as obvious a conclusion as it might seem. Having the security screen always display one message or another runs a risk of user fatigue in the long term. A longer term study needs to be done to determine whether affirmative displays of insecure modes continue to result in increased speed and accuracy over the long term. However, the short term data in this study is consistent with the idea that explicit notification helps users reach faster determinations.

These findings provide some hope that while security interfaces have often fallen short of providing the tools users need to effectively make good security decisions, better interfaces are possible. Accuracy rates increase significantly when interfaces avoid specific problems. Many existing systems have not been designed to avoid these mistakes, so continued poor results for these systems should not come as a surprise. With further study, security interfaces may one day achieve significantly higher levels of accuracy at reduced cognitive overhead.

5 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1018928. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.