

User Studies on Security: Good vs. Perfect

Volker Roth

FX Palo Alto Laboratory

Thea Turner

FX Palo Alto Laboratory

1 Introduction

Increasingly, our daily activities are supported by networked computing. Coincident with this trend, criminals have been exploiting the vulnerabilities of networks to cause all sorts of mayhem, resulting in harm to people's privacy and property. It is ironic that some of this damage is caused by people's own inattention to secure practices.

Understanding the behaviors that expose people to risks should result in better security systems and practices. Furthermore, research in the area of usable security has explored user interface mechanisms to reduce user susceptibility to various forms of attacks and to improve users' ability to interact with protection mechanisms.

Several studies report some success at reducing risky behavior in the laboratory setting [1, 7, 6]. Our personal experience in this area includes the investigation of the usability and security of novel security mechanisms:

- A human-assisted mechanism through which two mobile devices can establish a secure ad hoc wireless connection [5]
- A mechanism to enter one's PIN into a terminal so that a shoulder surfer does not learn it even if all input and output can be observed [3, 2]
- Electronic mail that is secured in a best-effort fashion and that would be easier to use than classic approaches [4]
- Use of a prototype video security system by police officers to handle a simulated airport security breach (unpublished study performed at Motorola Labs).

Yet, we believe that many users would fail to protect themselves in spite of the enhanced security and even if studies suggest that users would accept a novel mechanism and would perform well with it.

This can in part be explained by well-known artifacts of staged experiments such as demand characteristics (subjects try to please the experimenter), task focus (subjects do not question assigned tasks but are eager to succeed in them) and obedience to authority.

However, studies of subjects' responses to threat and their interaction with novel security mechanisms introduce additional complexities. The threats and environmental conditions that exist in the real world go beyond what can be recreated in a laboratory setting in an ethical and responsible fashion. This limits our ability to study security aspects of human computer interaction and also limits that validity of conclusions we draw from experiments.

We take a closer look at the nature of the limitations in §2. Then we discuss how some of these limitations can be attenuated. A study does not have to generalize entirely

to provide useful insights into how the usability and security of networked systems can be improved.

2 Limitations

Ethical considerations limit our ability to truly study user behavior in the face of threats to post-mortem studies of real cases or experimental setups with a very low level of personal risk for subjects. Even at low levels of risk, the experimenter may face legal and ethical risks. If we threatened subjects (with or without their awareness) in the way ruthless criminals do, then we deserve and must accept the same punishment. Few researchers would want to bear the consequences *and* to obtain results by such unethical means.

There would still be ethical concerns if we obtained users' informed consent to participate in a study where we observed their responses to actual threats. Additionally, subjects' behavior may not be authentic because: they anticipate a threat and are alert; they do not truly believe that the threat is real; they would expect to be fully compensated for the results of any attacks, which limits the threat. If none of the cases above holds and the subject nevertheless participates in the study, then the subject is not behaving rationally, which may not be representative of typical user behavior.

It is difficult to construct an experimental design that successfully mimics the real world environment in all significant ways and we cannot know whether the constraints in the laboratory task differ in meaningful ways from the actual task's constraints.

We faced these difficulties in one of our recent studies. We were interested in how much work subjects were willing to spend on a task that serves to establish secure wireless communication between devices. We offered qualitative advice on three different levels of security. Afterwards, we asked subjects to estimate the amount of work they had done. To our surprise, subjects underestimated their amount of work. Would they have expended as much effort if they were doing the task in a real world setting without the demand characteristics of the experiment? Their concern for the privacy of their personal documents may have resulted in more effort spent on security. Or they may be in a hurry and unwilling to take more than the minimum amount of time necessary for security against a nebulous, low probability threat.

While we *cannot* conclude from this that users would spend *any* work on the task in a real setting, we *can* assume that, *if* users perform the task, then they may invest more work than they thought they had.

3 Strategies to Overcome Some Limitations

In the studies we conducted, we had to make compromises between what we wanted to investigate and what we felt was feasible with the available resources. For instance, we asked subjects not to use their actual PIN numbers in PIN entry tasks, but rather to use a different number they would remember easily. This instruction was meant to prevent subjects from disclosing confidential information and also to protect us against future allegations of misuse of that information.

At the same time, we accepted that subjects' retrieval performance would not be accurate when compared to a real setting since they did not have time to internalize the chosen numbers as much as they had their actual PINs.

Training with fake credentials With more resources on our hands, we could have attempted to train subjects with chosen PIN numbers until we observe a flattening of the learning curve. However, all of our subjects were unpaid volunteers and therefore we refrained from increasing their load in our experiments. Nevertheless, training subjects with fake credentials appears to be one step to improve the accuracy of comparisons.

It may also be feasible to perform studies with a small trusting group of subjects using actual and fake credentials in an attempt to quantify the impact that the use of false credentials in an experiment may have on the results. Perhaps this would allow us to establish a "correction factor" to experimental results obtained in a larger study using fake credentials.

User-verifiable assurance A second option would be to build an entry terminal that provides *user-verifiable assurance* that no PIN numbers are leaked. Such a device would have to be based on simple mechanics and wiring that can be scrutinized and reconfigured by the subject. Consider a PIN pad where keys are wired to a plug-board. A reader is wired to a plug-board as well. Subjects are instructed to connect the boards using plugs while the experimenter is absent, which introduces a random permutation to the entered PIN. The device can be build of Plexiglas so that no additional electronics can be hidden in it. During the experimentation, subjects cover the plug-board settings with a black cloth. In this fashion, subjects are assured that they can use personal PIN numbers without risk of disclosure, and experimenters can repudiate false allegations if one of the subjects becomes the victim of a real attack.

While this example is useful only in a very specific setting the general idea might be applicable in other situations as well, which means that there is an opportunity to innovate in the area of experiment design and mechanism support for studies of usable security.

Comparative studies One strategy for studies on usable security is to construct experimental designs that compare either several different mechanisms, levels or versions of the same mechanism or multiple scenarios with different salient characteristics. Instead of trying to quantify the usability and security of a mechanism directly, we should be able to assess the *relative* behavior and performance of subjects in comparable tasks. Differences that can be observed in an experiment should have a similar effect in real settings, assuming we take care not to introduce environmental or contextual biases. We can draw conclusions about the relative merits of the mechanisms tested by examining their relative ranks.

For example, Uzun et al. [7] examined mechanisms for secure pairing between two devices. Had they used a single method, it would be difficult to know how successful it would be in practice. By comparing performance on several mechanisms, they were able to show that the least usable one was also the most secure. By examining performance on the three methods, they made adjustments that resulted in a mechanism that was both usable and secure.

Another method to facilitate comparative studies is to make materials available online that can be used to replicate the studies with different (or even the same) mechanisms [8].

4 Conclusions

The impracticality and ethical concerns of replicating accurate threat scenarios in a laboratory limit our ability to generalize quantitative results to realistic settings. Comparative data is more likely to lead to usable security through the evaluation of user behavior under different conditions. The ability to compare results across researchers (or labs) will also allow different mechanisms to be directly compared.

Innovations that increase the similarity between the experimental conditions and real-world settings will also increase our ability to generalize results to realistic situations. Furthermore, targeted study of differences between laboratory settings and real settings using very focused trusting groups allow us not to obtain general results but to understand the potential “correction factors” to laboratory results.

References

- [1] Goodrich, M. T., Sirivianos, M., Solis, J., Tsudik, G., and Uzun, E. Loud And Clear: Human verifiable authentication based on audio. In *Proc. 26th International Conference on Distributed Computing Systems* (July 2006), IEEE.
- [2] Volker Roth and Kai Richter. How to fend off shoulder surfers. *Journal of Banking and Finance*, 30(6):1727–1751, June 2006. Article in Press, doi:10.1016/j.jbankfin.2005.09.010.
- [3] Volker Roth, Kai Richter, and Rene Freidinger. A PIN entry method resilient against shoulder surfing. In *Proc. 11th ACM Conference on Computer and Communications Security*, Washington, DC, USA, October 2004.
- [4] Volker Roth, Tobias Straub, and Kai Richter. Security and usability engineering with particular attention to electronic mail. *International Journal of Human-Computer Studies*, 63:51–73, July 2005. Special Issue HCI research in privacy and security.
- [5] Volker Roth, Wolfgang Polak, and Eleanor Rieffel. A Ring to Rule Them All, Secure and Usable Pairing With a Human Sampling Function. February 2007. Under submission.
- [6] Desney S. Tan and Pedram Keyani and Mary Czerwinski. Spy-resistant keyboard: more secure password entry on public touch screen displays. In *Proc. 19th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction*, 1–10, 2005.
- [7] Uzun, E., Karvonen, K., and Asokan, N. Usability analysis of secure pairing methods. In *Proc. Usable Security Workshop (USEC)* (Lowlands, Scarborough, Trinidad/Tobago, Feb. 2007). Co-located with 11th Conference on Financial Cryptography and Data Security.
- [8] Various authors. Security User Studies Construction Kit. <http://cups.cs.cmu.edu/soups/2006/workshop-kits/kits.html>