Standard Metrics and Scenarios for Usable Authentication

Scott Ruoti
Brigham Young University
scott.ruoti@isrl.byu.edu

Kent Seamons
Brigham Young University
seamons@cs.byu.edu

1. INTRODUCTION

There is a constant flow of new authentication schemes proposed in the literature. In the past, most proposed schemes were not evaluated empirically, though in recent years there has been an increase in the number of authentication systems that have undergone a user study. Still, most of these user studies employ ad-hoc metrics (e.g., task completion time) and a unique scenario. Bonneau et al. [2] included usability criteria in their heuristic evaluation of various types of web authentication mechanisms.

The use of ad hoc and disparate metrics and scenarios makes it difficult to compare the relative merit of these various proposals. This produces disjointed results that hinder our ability to make more rapid, scientific progress toward usable authentication systems. Based on our experience, we believe that the community would benefit significantly from the adoption of standard metrics and scenarios for use in the empirical evaluation of authentication schemes.

2. DIRECT COMPARISON

Adoption of standard metrics and scenarios in usability studies of authentication systems would allow the results to be directly compared. This would bring value to the community by allowing us to determine whether we are making systematic progress towards more usable and secure authentication, or whether we are all just "spinning our wheels."

Our opinions are based on our study comparing the usability of seven web authentication systems [9]. The systems span three categories: federated single sign-on (Google OAuth 2.0, Facebook Connect, Mozilla Personas), email-based identification and authentication [6] (SAW [11], Hatchet [9]), and QR-code-based (WebTicket [7], Snap2Pass [5]).

Our usability studies were organized in a tournament structure. In the preliminary round, we conducted within-subject tests for all the systems in a single category (e.g., federated). The winner from each category advanced to a championship round where we conducted a within-subjects usability test of three heterogeneous systems. To our knowledge, this was the first user study to compare a heterogeneous collection of web-authentication systems.

Evaluating each system using common metrics and scenarios provided a stronger basis for directly comparing their usability. Our results showed that users prefer federated, single sign-on and Snap2Pass. Our results also demonstrated that several authentication proposals (SAW, Hatchet, WebTicket) were rated as less usable than existing password-based authentication schemes. Interestingly, WebTicket had previously been evaluated with a user study, and our study found the same usability benefits and pitfalls. What our study added was an understanding of where WebTicket fit in the greater ecosystem of authentication schemes; precisely the benefit provided by comparing systems in a standardized manner.

2.1 Standard Metrics

Some metrics used in prior studies include task completion time, error rates, and recall rates. We promote the use of the System Usability Scale (SUS) as a standard metric for calculating the relative usability of authentication schemes. SUS [3, 4] is a standard metric from the usability literature, and has been used in hundreds of usability studies [1]. Our experience has also shown that a system's SUS score is consistent across different sets of users [8]. Moreover, Tullis and Stetson compare SUS to four other usability metrics (three standard metrics from the usability literature and their own proprietary measure) and determined that SUS gives the most reliable results [10].

The SUS metric is a single numeric score from 0, the least usable, to 100, the most usable, that provides a rough estimate of a system's overall usability. To calculate a system's SUS score, participants first interact with the system and then answer ten questions relating to their experience . Answers are given using a five-point Likert scale (*strongly agree* to *strongly disagree*). The questions alternate between positive and negative statements about the system being tested. Participants' answers are assigned a scalar value and then summed to produce the overall SUS score, and the system with the highest average SUS score is the most usable.

SUS produces a numeric score for a non-numeric measure (i.e., usability), making it difficult to intuitively understand how usable a system is based solely on its SUS score. As part of an empirical evaluation of SUS, Bangor et al. [1] reviewed SUS evaluations of 206 different systems and compared these scores against objective measurements of the various systems' success in order to derive adjective-based

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

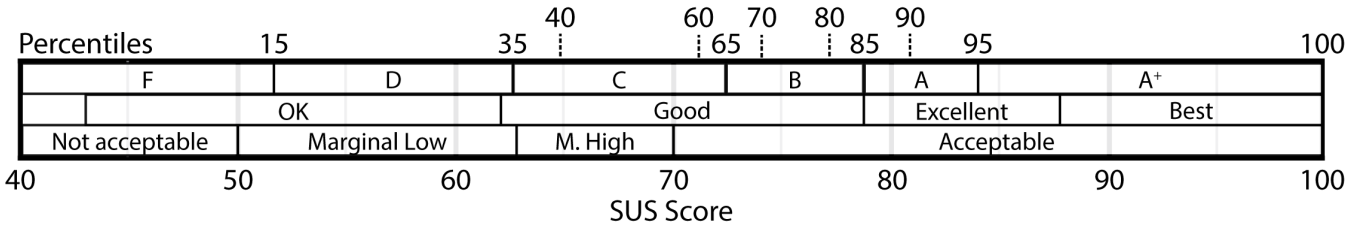

Figure 1: Adjective-based ratings to help interpret SUS scores

ratings for SUS scores. We have summarized these ratings and their correlation to SUS scores in Figure 1.

In our study, we found that SUS was a strong predictor of participants' preferred authentication schemes, and was consistent across multiple studies. As such, SUS is a promising standard metric for comparing authentication schemes. However, SUS was not designed for usable security. One potential direction is to extend SUS with questions targeted toward usable security. At the workshop, we are interested in comparing our experiences to others that have used standard metrics, and also discussing the pros and cons of the various standard metrics that are available.

2.2 Standard Scenarios

As part of our web authentication user studies, we built two websites: a **forum** website where users could get help with smartphones,[1] and a **bank** website.[2] We chose these two types of websites because they represented diametrically different information assurance needs. We then tested each of the seven systems in the context of these websites. The reason for doing this was to limit the number of confounding factors. The only differences between the systems were due to authentication system details and not the application.

We have made the source code for these websites publicly available[3] so that others can re-use them to test other authentication systems. The sharing of implementations of compelling use case scenarios can reduce the start-up costs to run usability studies for new authentication system proposals. We are interested in identifying alternative scenarios for web-based authentication, as well as compelling scenarios for emerging forms of authentication.

3. RECOMMENDATIONS

Based on our experience, we propose that usability studies incorporate standard metrics like SUS for vetting all new authentication proposals. In the case of SUS, we recommend that new authentication systems should exceed a baseline SUS score of 68 before receiving serious consideration.[4] This will provide a basis to make cross-system comparisons and establish a minimum threshold for vetting new proposals. Any new system proposal that fails to achieve a sufficiently high average usability rating is very unlikely to see widespread adoption.

We should also design standard scenarios and make implementations available to researchers to reduce the effort to conduct usability studies; such studies will have a stronger basis for comparison across systems. Widespread adoption of these recommendations has the potential to significantly enhance our effort as a community to identify and focus on authentication systems that have the strongest potential to be usable.

[1] `https://forums.isrl.byu.edu`

[2] `https://bank.isrl.byu.edu`

[3] `https://bitbucket.org/isrlauth/battle-website`

[4] This is based on the SUS scores of the systems from our study.

4. REFERENCES

[1] A. Bangor, P. Kortum, and J. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3), 2009.

[2] J. Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Symposium on Security and Privacy*. IEEE, 2012.

[3] J. Brooke. SUS — a quick and dirty usability scale. In *Usability Evaluation in Industry*. CRC Press, 1996.

[4] J. Brooke. SUS: A retrospective. *Journal of Usability Studies*, 8(2):29–40, 2013.

[5] B. Dodson, D. Sengupta, D. Boneh, and M. S. Lam. Secure, consumer-friendly web authentication and payments with a phone. In *International Conference on Mobile Computing, Applications, and Services*. Springer, 2012.

[6] S. L. Garfinkel. Email-based identification and authentication: An alternative to pki? *IEEE Security & Privacy*, (6):20–26, 2003.

[7] E. Hayashi, B. Pendleton, F. Ozenc, and J. Hong. Webticket: Account management using printable tokens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012.

[8] S. Ruoti, N. Kim, B. Burgon, T. Van Der Horst, and K. Seamons. Confused Johnny: When automatic encryption leads to confusion and mistakes. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS)*. ACM, 2013.

[9] S. Ruoti, B. Roberts, and K. Seamons. Authentication melee: A usability analysis of seven web authentication systems. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015.

[10] T. S. Tullis and J. N. Stetson. A comparison of questionnaires for assessing website usability. In *Usability Professional Association Conference*, 2004.

[11] T. W. Van Der Horst and K. E. Seamons. Simple authentication for the web. In *Third International Conference on Security and Privacy in Communication Networks (SecureComm)*. IEEE, 2007.