Authentication Melee: A Usability Analysis of Seven Web Authentication Systems

Scott Ruoti, Brent Roberts, Kent Seamons
Internet Security Research Lab
Brigham Young University
Provo, UT 84602
{ruoti, roberts}@isrl.byu.edu, seamons@cs.byu.edu

ABSTRACT

Passwords continue to dominate the authentication landscape in spite of numerous proposals to replace them. Even though usability is a key factor in replacing passwords, very few alternatives have been subjected to formal usability studies, and even fewer have been analyzed using a standard metric. We report the results of four within-subjects usability studies for seven web authentication systems. These systems span federated, smartphone, paper tokens, and email-based approaches. Our results indicate that participants prefer single sign-on systems. We report several insightful findings based on participants' qualitative responses: (1) transparency increases usability but also leads to confusion and a lack of trust, (2) participants prefer single sign-on but wish to augment it with site-specific low-entropy passwords, and (3) participants are intrigued by biometrics and phone-based authentication. We utilize the Systems Usability Scale (SUS) as a standard metric for empirical analysis and find that it produces reliable, replicable results. SUS proves to be an accurate measure of baseline usability. We recommend that new authentication systems be formally evaluated for usability using SUS, and should meet a minimum acceptable SUS score before receiving serious consideration.

Categories and Subject Descriptors

K.6.5 [**Management Of Computing And Information Systems**]: Security and Protection—*Authentication*; H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces—*Evaluation/methodology*

Keywords

Authentication; security; usability; System Usability Scale

1. INTRODUCTION

Passwords continue to dominate the authentication landscape. Bonneau et al. [6] analyzed a broad collection of web authentication schemes designed to replace passwords. They demonstrated that passwords have a unique combination of usability, security, and deployability that has proven difficult to supplant. While Federated identity systems (i.e., Google OAuth 2.0, Facebook Connect) and password managers (e.g., LastPass) are seeing some success, these systems are designed to enhance passwords usage rather than disrupt it.

While Bonneau et al. presented a heuristic-based approach for evaluating the usability of authentication schemes, it is also imperative that authentication systems are subjected to empirical usability analysis. We survey the publications cited by Bonneau et al. and discover that only four of the twenty-three publications report the results of an empirical usability study [9, 15, 24, 29]. Moreover, Chiasson et al. [9] is the only study that compares its proposed system against another competing authentication system. Most troubling, none of the systems are analyzed using a standard usability metric, making it impossible to determine which of the four systems has the best usability. Without a standard metric, there is no means by which a new proposal can be evaluated to determine whether it has better usability than existing systems.

In this paper, we report the results of a series of within-subjects empirical usability studies for seven web authentication systems. The seven authentication systems are heterogeneous and span federated, smartphone, paper token, and email-based approaches. Our studies are the first to compare a heterogeneous collection of web authentication proposals. Our research goals are two fold:

1. *Determine which system has the best overall usability.* This is accomplished using the System Usability Scale (SUS) [7, 8], a standard usability metric that has been used in hundreds of studies [2, 3].

2. *Explore which authentication features users prefer and which features they dislike.* In our studies, participants use multiple authentication systems and provide feedback describing what they like and what they would change.

In the results of our studies, federated and smartphone-based single sign-on receive the best overall usability ratings. We also report insightful information from participants' qualitative responses. We find that systems with minimal user interaction are rated as highly usable, but are also described by participants as confusing and unworthy of trust. Although participants rate the usability of single sign-on highly, they are interested in augmenting it with additional low-entropy passwords. Our results also show that

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741683.

over half of participants are willing to use new authentication systems in their everyday life but they are most interested in adopting systems they perceive as different and innovative (e.g., biometrics, phone-based authentication).

Finally, our results validate SUS as an appropriate metric for comparing the usability of authentication systems because the SUS score for a given system is consistent across different participant groups and proves to be a strong indicator of users' preferences. We recommend that all new authentication proposals be formally evaluated for usability with SUS and that no proposal should receive serious consideration until it achieves a minimum acceptable SUS score.

2. AUTHENTICATION TOURNAMENT

In order to attain widespread deployment it is essential that new authentication systems not only be more secure than passwords, but must also provide tangible usability benefits that incentivize adoption. Very few web authentication systems have been evaluated using an empirical study. Fewer still have been analyzed using a standard usability metric or compared to alternative authentication systems. This makes it impossible to determine which of the existing systems is most usable.

As a first step to answering this question, we conduct empirical usability studies on seven web authentication systems. We use the System Usability Scale to determine which system is most usable. Also, we structure our usability studies as a tournament to gather qualitative data from participants regarding which authentication features are most important to them.

2.1 System Usability Scale

To address our first research goal, *which system has the best overall usability*, we measure each systems usability based on a standard usability metric. The System Usability Scale (SUS) [7, 8] is a standard metric from the usability literature that we adopt as part of our methodology. SUS has been used in hundreds of usability studies [3] and the original SUS paper [7] has been cited over 2,450 times.[1] Our prior work has also shown that a system's SUS score is consistent across different sets of users [20]. 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 [27].

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 (see Table 1). 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 (see Table 2) 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. [3] 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 ratings for SUS scores. These ratings and their correlation to SUS scores are given in Figure 1. We report these adjective-based ratings along with SUS scores to provide readers with a better intuition of each system's usability.

[1]Citation count retrieved from Google Scholar on Feb 28, 2015.

1)	I think that I would like to use this system frequently.
2)	I found the system unnecessarily complex.
3)	I thought the system was easy to use.
4)	I think that I would need the support of a technical person to be able to use this system.
5)	I found the various functions in this system were well integrated.
6)	I thought there was too much inconsistency in this system.
7)	I would imagine that most people would learn to use this system very quickly.
8)	I found the system very cumbersome to use.
9)	I felt very confident using the system.
10)	I needed to learn a lot of things before I could get going with this system.

Table 1: The ten SUS questions

	Questions 1,3,5,7,9	Questions 2,4,6,8,10
Strongly Agree	10	0
Agree	7.5	2.5
Neither Agree or Disagree	5	5
Disagree	2.5	7.5
Strongly Disagree	0	10

Table 2: SUS score card

2.2 Tournament Structure

To address our second research goal, *which features of authentication do users prefer and which do they dislike*, we have participants use multiple authentication systems and then have them provide feedback on their experience. We believe that after participants have used multiple systems that they will be better able to articulate their opinions on authentication. One option would be to perform a full combinatorial comparison, but this would be prohibitive in terms of time and cost. For example, if each system is tested by 20 participants,[2] and an individual participant tests systems, it would require $\binom{7}{2} * 20 = 21 * 20 = 420$ participants, 27 person-days of effort, and $4,200 USD to complete the study.[3] Alternatively, having each participant using all the authentication systems could result in study fatigue that would bias the results.

Instead, we model our study after a tournament bracket. We first arrange the seven web authentication systems into three groups based on common features. These groups are federated single sign-on, email-based, and QR code-based. For each of the groups we conduct a separate usability study, and the system with the highest SUS score in each study is selected as a winner. The three winners are then compared

[2]20 participants is an average sample size used in security usability studies and is enough for statistical significance.

[3]The calculation assumes a 45-minute study at $10 per user. These costs grow factorially in the number of systems tested.

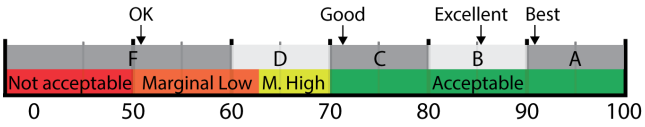

Figure 1: Adjective-based ratings to help interpret SUS scores

to each other in a championship round usability study. This methodology allows us to gather qualitative user feedback from participants who have tested similar systems and also participants who have tested dissimilar systems.

The breakdown of systems into the tournament bracket is given in Figure 2 and the remainder of this section describes the contestants in our authentication tournament.

2.2.1 Federated Single Sign-on

In federated single sign-on, all authentication responsibility is centralized in a single identity provider (IDP). Instead of websites maintaining their own collection of usernames and passwords, websites instead rely on the IDP to verify the identity of users visiting their website. The IDP is free to use whatever method it wants to authenticate users, though the three systems in our tournament all use usernames and passwords.

We select three federated single sign-on systems for inclusion in our tournament: Google OAuth 2.0, Facebook Connect, and Mozilla Persona. Google OAuth 2.0 and Facebook Connect are chosen because they are the only authentication systems other than current password-based authentication that are widely adopted. Since both Google and Facebook store personal information for users, it is possible that users might reject both systems for fear that their personal information will be leaked [18]. To address this concern, we also include Mozilla Persona, a federated single sign-on system that does not store users' personal information.

2.2.2 Email-based Single Sign-on

Email-based single sign-on is similar to federated single sign-on, but instead of centralizing authentication responsibilities into a single entity (e.g., Google, Facebook), they are instead delegated to email providers [13]. Users prove their identity by demonstrating their ability to either send or receive email. The advantage over federated single sign-on is that users have the freedom to choose which email providers they trust to be an identity provider.

We select two systems for this group: Simple Authentication for the Web (SAW) [28] and Hatchet. SAW was developed in our research group and was cited for this category in the Bonneau et al. survey. SAW authenticates a user by sending them an email with a link they can click to log in to the website. To increase the security of authentication, SAW requires the user to click the link on the device they want to be authenticated on. Hatchet is a variant of SAW that we developed for the purpose of this study. Hatchet replaces the link sent in SAW with a one-time password (OTP). This OTP is then entered into the website the user is logging in to.[4] Unlike SAW, Hatchet allows users to retrieve email on one device and be authenticated on another device.

2.2.3 QR Code-based

For our last group, we select the two most recent authentication proposals we are aware of: WebTicket [14] and Snap2Pass [12]. Both of these systems use QR codes and require a physical token to authenticate the user: a piece of paper and a smartphone respectively. In WebTicket, a user's credentials are encoded in a QR code that is printed and stored by the user (their WebTicket). The user authenticates to the website by scanning their WebTicket using their computer's webcam. WebTicket was originally presented as a browser plugin, but we have modified it to allow websites to deploy WebTicket for authentication. We believe that this is a more likely deployment scenario, as users have proven to be reticent to install browser plugins [20, 18].

Snap2Pass is a single sign-on authentication system where the user's phone acts as an IDP. The user first pairs their phone with the website by using the Snap2Pass application to scan a QR code provided by the website. Later, when the user authenticates to the website, they are presented with another QR code to scan. After scanning this QR code, the user's phones will verify the identity of the user to the website and the user is logged in.

3. METHODOLOGY

During the summer and fall of 2014, we conducted four IRB-approved studies analyzing the usability of seven web authentication systems. The studies varied as to which authentication systems were tested, but otherwise the content of the studies remained constant. This section gives an overview of the studies and describes the task design, study questionnaire, study development, and limitations.

3.1 Study Setup

The four studies were conducted between June and October 2014: June 24–July 12, July 28–August 23, October 7–October 11, and October 13–October 24. The first three studies evaluated the federated, email-based, and QR code-based groups respectively, and the fourth study is the championship round usability study. In the first study (federated), participants were randomly assigned two of the three authentication systems in the group, and in the second (email-based) and third studies (QR code-based) participants were assigned to use both systems in the group. In the fourth study (championship round), participants were assigned all three systems.[5]

In total, 106 individuals participated in our studies: 24 participants in the first study, 20 participants in the second study, [6] 27 participants in the third study, and 35 participants in the fourth study. Each individual was allowed to

[4]This use of OTPs is not unique to Hatchet [1], but to our knowledge there is no authentication system which employees OTPs and can be used to authenticate to arbitrary websites.

[5] We modified the study to assign participants three systems for two reasons: (1) in the first three studies participants showed no signs of study fatigue after evaluating two authentication systems and (2) we were interested in the qualitative responses of participants who had been assigned three heterogeneous authentication systems.

[6]We are unsure why fewer students signed up for the second study, though we speculate that it might be due to the

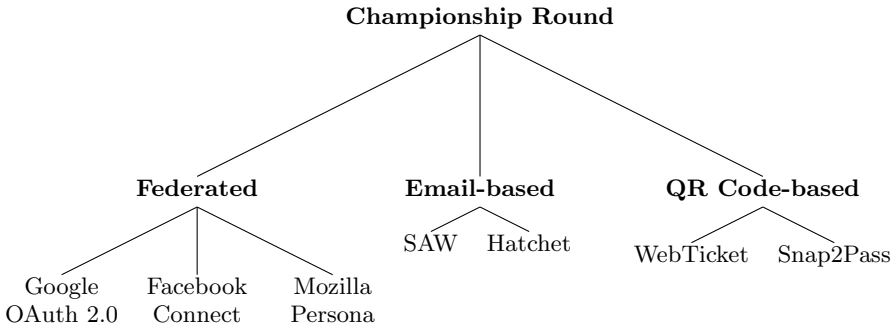

Figure 2: Authentication tournament bracket

participate in only one of the four studies. Participants took a minimum of 20 minutes and a maximum of 45 minutes to complete their study and were compensated $10 USD for their efforts. When using Snap2Pass, participants were provided with a Nexus 5 smartphone with the Snap2Pass application pre-installed. When using WebTicket, participants were provided with a black and white laser printer, a pair of scissors, and a 1080p webcam.

3.1.1 Quality Control

The results for eight participants are discarded for various reasons:

- Two participants, both in the second study (email-based), had the authentication emails generated by SAW marked as spam.[7] The survey coordinator was unable to resolve this problem and the participants were unable to complete the study.
- Three participants, one in the third study (QR code-based) and two in the fourth study (championship round), were non-native English speakers and were unable to understand the study's instructions.
- Three participants, one in the third study (QR code-based) and two in the fourth study (championship round), skipped a task and did not finish registering a necessary account. The study coordinator was unable to resolve this problem and the participants were unable to complete the study.

After removing results from these 8 participants we are left with results from 98 participants: 24 participants in the first study (federated), 18 in the second study (email-based), 25 in the third study (QR code-based), and 30 in the fourth study (championship round). The remainder of this paper will refer exclusively to these 98 participants.

3.1.2 Participant Demographics

We recruit participants for our study at Brigham Young University. All participants are affiliated with Brigham Young University,[8] with the overwhelming majority being undergraduate students: undergraduate students (93; 95%), graduate students (3; 3%), faculty (1; 1%), did not provide demographic information (1; 1%). Participants had a variety of majors, 51 in total, with the highest percentage studying exercise science (8 participants). No other major had more than five participants. We recruited broadly across campus to avoid attracting primarily technical majors. Participants were asked to self report their level of technical skill, with most reporting an intermediate level of knowledge. Table 3 contains a breakdown of participant demographics by study.

	Gender		Age		Technical Skill		
	Male	Female	18–24 years old	25–34 years old	Beginner	Intermediate	Advanced
Federated (n = 24)	58% 14	42% 10	83% 20	17% 4	13% 3	79% 19	8% 2
Email (n = 18)	67% 12	33% 6	78% 14	22% 4	28% 5	72% 13	0% 0
QR Code (n = 25)	52% 13	48% 12	88% 22	12% 3	12% 3	60% 15	28% 7
Championship (n = 30)[1]	67% 20	33% 10	77% 23	23% 7	13% 4	83% 25	4% 1
Total (n = 97)[1]	62% 59	38% 38	81% 79	29% 18	15% 15	75% 73	10% 9

[1] One participant in the QR code-based group did not provide demographic information, explaining the smaller number of participants reported in this table.

Table 3: Participant demographics

3.2 Task Design

We built two WordPress websites for the purpose of our studies: a **forum** website where users could get help with smartphones,[9] and a **bank** website.[10] We chose these two types of websites because they represented diametrically different information assurance needs. At a forum website there is little personal information stored, and so even if the user's account is stolen there is only minimal risk of harm. Conversely, users have been shown to be extremely cautious when it comes to online bank accounts [22]. Studying web-

fact that the majority of our participants were undergraduate students at Brigham Young University and finals for Summer term fell on August 13–14.

[7] We suspect that the emails were marked as spam because they contained both the words "bank" and "click on the link". Different wording could have avoided this problem.

[8] We did not require this affiliation.

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

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

sites with different information assurance needs allows us to examine whether users are amenable to a given authentication system being deployed to all websites, or only to websites that do not store personal information.

During the studies, participants are assigned two or three authentication systems in random order. For each authentication system, participants are given six tasks to complete (three for each website). For each task, participants are instructed on how to use the website to complete the task. Participants are not instructed on how to use any of the authentication systems, as one aspect of usability is how well an authentication system facilitates a novice user. Between each task, participants are logged out of both websites, ensuring that participants use the assigned authentication system for each task.

Below is a summary of the six tasks:

Task 1.
Participants create a new account at the **forum** website using the assigned authentication system.

Task 2.
Participants modify an existing **bank** account to allow login using the assigned authentication system.

Task 3.
Participants log into the **forum** website and create a post in the "New User" forum.

Task 4.
Participants log into the **bank** website and look up their checking account balance.

Task 5.
Participants log into the **forum** website and search for a specific post.

Task 6.
Participants log into the **bank** website and transfer money from one account to another.

3.2.1 Authentication System Implementation

For this study we implemented all seven authentication systems. We did this for two reasons: first, existing implementations of SAW, Hatchet, WebTicket and Snap2Pass are non-existent[11] and second, by implementing the systems ourselves we could assure a consistent user experience.

Source code for our implementations of these systems, as well as the **forum** and **bank** websites, is available at `https://bitbucket.org/isrlauth/battle-website`.

3.3 Study Questionnaire

We administer our study using Qualtrics' survey software. The survey begins with an introduction and a set of demographic questions.

Participants receive written instructions on how to complete the six study tasks for a particular authentication system. Participants use their own email and social network accounts to complete tasks. After completing the tasks, participants answer the ten SUS questions. Next, participants describe which features of the assigned authentication system they enjoy and which they would change. Lastly, participants indicate whether they would prefer to use the assigned authentication system over current password-based authentication and why. This process is then repeated for each assigned authentication system.

At the end of the survey, participants were asked several final questions. First, participants were asked what their favorite authentication system was: whether it was one of the systems they tested or current password-based authentication. They were also asked to explain why the selected system was their favorite. Lastly, participants were asked to describe their ideal authentication system. While most participants are not software engineers or user experience designers, we believe that asking this question serves two purposes: (1) it allows participants to synthesize all the systems they have used and extract what they consider the best from each and (2) it allows participants to mention authentication features that excite them but are not a part of any of the assigned systems.

3.4 Survey Development

After implementing the federated single sign-on systems, we developed the study tasks and questionnaire. We then had a convenience sample of nine individuals from our research institute complete the study. Based on their feedback we made some alterations to wording of the task instructions. After making these changes we began the first usability study (federated).

During this first study (federated), we noticed that a small number of participants were confused about how to complete the second task. In each case, the study coordinator was able to explain to them where to go on the **bank** website to complete the task and we did not need to discard any of the participant's responses. To avoid having participants ask the study coordinator for assistance in the three remaining studies we made a slight visual modification to the **bank** website. This change was universal for all the authentication systems and did not affect their functionality.

During the second usability study (email-based), Gmail began marking some of the authentication emails as spam. To our knowledge, four participants encountered this problem. This problem prevented the first two participants from completing the study and their results were discarded. For the latter two participants, the study coordinator was able to diagnose the problem and help them complete the study. In the fourth study, which once again included SAW, we added a note to the **bank** tasks to indicate to participants that this might occur and how to remedy the problem.

3.5 Limitations

While our studies included students with a diverse set of majors and technical expertise, it would be beneficial for future studies to test authentication systems using a non-student population. It is likely that a large number of participants are already familiar with Google OAuth 2.0 and Facebook Connect and this may have affected their opinions. A laboratory study only captures initial impressions. Participants might feel very differently after day-to-day use or discover new problems. Also, we only study seven authentication systems, which permits us to classify the usability of only a small fraction of existing authentication proposals. Future research could examine additional authentica-

[11] We contacted the authors of WebTicket and Snap2Pass and requested their implementations, but we received no reply.

		SUS			Better than Passwords		Is Participants Favorite System
		Mean	Standard Deviation	Median	Yes	Maybe	
n=16	**Google**	**72.0**	12.4	72.5	**31%**	38%	**31%**
	Facebook	71.4	13.5	72.5	13%	31%	25%
	Mozilla	71.8	10.8	71.3	**31%**	44%	25%
n=18	**SAW**	**61.0**	17.5	62.5	**28%**	28%	**44%**
	Hatchet	53.5	16.4	52.5	22%	44%	17%
n=25	WebTicket	57.9	16.9	60	20%	28%	4%
	Snap2Pass	**75.7**	17.8	82.5	**36%**	40%	**76%**
n=31	**Google**[1]	**75.0**	14.8	77.5	**26%**	32%	**29%**
	SAW[1]	53.2	16.2	55	6%	29%	0%
	Snap2Pass[1]	68.4	16.7	70	**26%**	39%	**29%**

The best performing system and metric for each usability study is given in **bold**. For the second, third, and fourth studies, participants used all available authentication systems and so $100\% - \Sigma(Favorite\ System)$ gives the percent of participants who preferred current password-based authentication to any of the assigned authentication systems.

[1] Championship round.

Table 4: SUS scores and participant preferences

tion systems in order to increase knowledge on the usability of authentication systems and help determine which systems are best-in-class and which system has the best overall usability.

4. RESULTS

In this section we report the quantitative results we gathered. Table 4 gives the SUS scores from the four usability studies and summarizes participants' authentication system preferences. Table 5 records whether the difference in the systems' SUS scores is statistically significant. Finally, Table 6 reports the mean time to authenticate for each system. The complete data set from the user study, including anonymous participant responses, is available in Ruoti's Master's Thesis [19].

The remainder of this section breaks down the individual results for each of the four usability studies. As mentioned in Section 2.1, in order to help the reader interpret the meaning of the SUS scores, we also report where these scores fall on Bangor's adjective-based scale [2, 3].

4.1 First Study – Federated

The SUS scores for Google OAuth 2.0, Facebook Connect, and Mozilla Persona were between 71 and 72, and the difference is not statistically significant. On Bangor's scale, all three systems are equal to or slightly above the "good" label, classified as acceptable, and receive a C grade.

Both Facebook Connect and Google OAuth 2.0 had similar registration and authentication times. In contrast, Mozilla Persona's registration and authentication times were two and four times greater, respectively. Even though there was a clear difference in mean time to authenticate, participants never mention this difference in their qualitative responses.

In deciding which authentication system they prefer, participants list trust in the federating party (i.e., Google, Facebook, Mozilla) as a key component. Many participants are hesitant to use Facebook Connect for fear that their social networking data would also be given to the website. Similarly, some participants are concerned that using Google OAuth 2.0 might increase the likelihood of their e-mail being hacked. There is little worry about Mozilla Persona in this regard.

According to our methodology, the winner of each usability study was decided based on highest SUS score. Since the difference of all three systems' SUS scores is not statistically significant, we attempt to break this tie based on which system has the highest number of participants who rate it as their favorite system. Once again, we find that all three systems perform similarly (Google – six participants, Facebook – five participants, Mozilla – five participants), and so we declare all three systems as winners. We still need a single system to move forward in the tournament and so we select Google OAuth 2.0, which had both the highest SUS score and the highest number of participants who rated it as their favorite system.

4.2 Second Study – Email-based

SAW's SUS score was higher than Hatchet's SUS score and this difference was statistically significant. As such, SAW is the winner of this round. Still, SAW's usability is not impressive. According to Bangor's scale, SAW's SUS score of 61 falls equidistant between the "excellent" and "good" label, is classified as having low-marginal acceptability, and given a D grade. Hatchet is slightly above the "OK" label, is classified as having low-marginal acceptability, and is given a failing grade.

While SAW was clearly the SUS champion in this category, participants using Hatchet and SAW took roughly equal amounts of time to register and authenticate (differences not statistically significant: registration — $p = 0.46$, authentication — $p = .27$).

4.3 Third Study – QR Code-based

Snap2Pass was the clear winner of this group, with a SUS score 17.8 points higher than WebTicket's SUS score (this difference was statistically significant). Additionally, only one participant indicated they would prefer WebTicket to Snap2Pass. According to Bangor's scale, Snap2Pass is slightly above the "good" label, is classified as acceptable, and receives a C grade. In contrast, WebTicket is between "OK" and "Good" (closer to "OK"), is classified as having low-marginal acceptability, and receives a D grade.

Participants' qualitative responses indicate that they felt both systems were fast, though comments made after the study indicate that they felt Snap2Pass was the faster of the two systems. These comments match the observations of the study coordinator who observed a significant number of participants struggle to authenticate quickly with WebTicket.

Two statistics in this study (QR code-based) vary significantly from the statistics in the other three usability studies. First, the median SUS score for Snap2Pass is significantly higher than its mean SUS score, indicating that their are several outliers who rate Snap2Pass very negatively, pulling its average down. In all the other results, including the fourth study when Snap2Pass is evaluated a second time, SUS scores are normally distributed. Second, 76% of participants in this study indicated that they are willing to replace current password-based authentication with Snap2Pass. In

	Google	Facebook	Mozilla	SAW	Hatchet	WebTicket	Snap2Pass	Google[1]	SAW[1]	Snap2Pass[1]
Google	—	.89	.94	.04	<.01	<.01	.47	.50	<.01	.45
Facebook	.89	—	.94	.06	<.01	.01	.42	.42	<.01	.54
Mozilla	.94	.94	—	.04	<.01	<.01	.43	.43	<.01	.47
SAW	.04	.06	.04	—	.05	.57	.01	<.01	.12	.15
Hatchet	<.01	<.01	<.01	<.05	—	.40	<.01	<.01	.96	<.01
WebTicket	<.01	.01	<.01	.57	.40	—	<.01	<.01	.30	<.01
Snap2Pass	.47	.42	.43	.01	<.01	<.01	—	.87	<.01	.12
Google[1]	.50	.42	.43	<.01	<.01	<.01	.87	—	<.01	.08
SAW[1]	<.01	<.01	<.01	.12	.96	.30	<.01	<.01	—	<.01
Snap2Pass[1]	.45	.54	.47	.15	<.01	<.01	.12	.08	<.01	—

2-tailed t-test. The participants for the second, third, and fourth study used all available authentication systems and within these groups statistical significance is calculated using the same population, while other significance values are calculated using equal variance. Only statistically significant results at the $p = .05$ level are shaded.

Row scheme scored higher than column scheme

Row scheme scored worse than column scheme

[1] Championship round.

Table 5: Comparison of system SUS scores

		Registration			Authentication				
		Task 1	Task 2	Average	Task 3	Task 4	Task 5	Task 6	Average
n=16	Google	46	43	44	3	10	2	2	**4**
	Facebook	53	23	**38**	7	6	3	4	**5**
	Mozilla	80	81	81	22	30	15	10	19
n=18	SAW	72	30	51	22	17	14	15	**17**
	Hatchet	51	29	**40**	27	20	19	17	21
n=31	Google[1]	51	38	**44**	3	2	2	2	**2**
	SAW[1]	55	34	**45**	62	42	17	25	36
	Snap2Pass[1]	76	-	76	13	14	13	11	13

All times are reported in seconds. We recorded participants' screens and use this data to calculate mean time to authenticate for all tasks except the second task of Snap2Pass. There are no results for the third study (QR code-based) due to video recording software failure.

[1] Championship round.

Table 6: Mean time to authenticate

the other three studies, only 60% of individuals indicated they were willing to replace current password-based authentication.

We are unsure as to what these anomalies mean, but report them in the interest of full disclosure. We are also unsure what caused these results, though we speculate it could be related to the fact that the second study had over a quarter of participants who rated themselves as having advanced technical skill (see Table 3).

4.4 Fourth Study – Championship Round

The championship round usability study consisted of the winners from the first three usability studies: Google OAuth 2.0, SAW, and Snap2Pass. The results are a tie between Google OAuth 2.0 and Snap2Pass, with SAW the clear loser. We apply the tie-break criteria from the first study (see Section 4.1), but the same number of participants chose Google OAuth 2.0 and Snap2Pass as their favorite system. For all three systems, there is no statistically significant difference between their scores in this study (championship round) and the previous three studies.

Since all three federated single sign-on systems tied in the first study, we declare federated single sign-on (collectively) and Snap2Pass to be the winners of our tournament.

5. DISCUSSION

In this section we begin with a discussion of SUS. We follow this with various insights gained from participants' qualitative responses. Finally, we report lessons learned while implementing the seven authentication systems.

5.1 System Usability Scale

SUS proves to be a highly reliable metric. SUS scores for Google OAuth 2.0, SAW, and Snap2Pass were consistent between the first three studies and the championship round study.[12] Within a single study, SUS scores for the systems are consistent regardless of the order in which participants use the systems, with all differences failing to be statistically significant.

Moreover, SUS is a good predictor of which system participants select as their favorite. In the first study (federated), all three federated systems had similar SUS scores, and an

[12]The differences in SUS scores is not statistically significant (see Table 5).

equal number of participants selected each of the three systems as their favorite. Likewise, in the second (email-based) and third (QR code-based) studies, when one system's SUS score was higher than the other system's SUS score, participants largely preferred the system with the higher SUS score. Most interesting, these preferences held between different sets of participants. The SUS scores for Google OAuth 2.0 and Snap2Pass are similar and the difference between the two is not statistically significant (see Table 5). This would indicate that an equal number of participants should prefer both systems, and this is indeed the case when they are evaluated in the championship round study (see Table 4).

While mean time to authenticate is reported in nearly every authentication usability study, our results indicate that mean time to authenticate is actually a poor measure of overall usability or participants' preferences. In the first study, Mozilla Persona had a much higher mean time to authenticate than either Google OAuth 2.0 or Facebook Connect, yet all three had similar SUS scores and were equally preferred by participants. Similarly, SAW and Hatchet did not differ significantly in mean time to authenticate, yet there was a clear distinction in both systems' SUS scores and participants' preferences.

Based on these results, we suggest that an empirical analysis using SUS be required for all future authentication system proposals. This allows new systems' SUS scores to be compared against existing proposals and validate whether these new proposals are improving upon the state-of-the-art. Additionally, we recommend that all new systems achieve a SUS score of 70 before they receive serious consideration. In our studies, only systems with a score of at least 70 (Google OAuth 2.0, Facebook Connect, Mozilla Persona, Snap2Pass) received consistently positive reviews from participants.

5.2 Transparency

Upon reviewing the results of the usability study (federated) we found that participants preferred systems that were transparent and required minimal interaction.[13] To verify that transparency improves usability, we administered a mini-study at the end of the second usability study (email-based). After completing the questionnaire for the second study, participants are then assigned a modified version of SAW. This modified version of SAW automates the process of retrieving and clicking links sent to user's email. Before beginning the six tasks, participants entered their email credentials into the new authentication system, and from then on whenever they click the login button they would immediately be logged into the website. Participants complete the same six tasks and answer the same questions as they did for all the other authentication systems.

The usability improvements of this modified version of SAW are striking. The modified version had a mean SUS score of 73.1, a standard deviation of 10.1, and median score of 75. This is an increase of 12.1 points over SAW's SUS score, and the difference is significant at the $p = 0.01$ significance level. This shows that transparency has a strong effect on perceived usability.

While these results demonstrate that transparency increases usability, transparency was not without its trade-offs. Minimal interaction with the authentication system prevents participants from understanding how the authentication system functioned and many participants have trouble trusting what they don't understand:

> *"I would like to understand more about how it works up-front. It doesn't feel secure."*
>
> *"If I understood how the system would prevent someone other than me from logging in I would use it."*
>
> *"I think it was very straightforward to use. Once again like with the other system, perhaps an explanation of how it protected information would give me more confidence in using it."*

This issue of transparency leading to confusion and lack of trust also appeared in our earlier research on secure webmail [20]. Future research could look closely at these trade-offs to discover what is an appropriate level of transparency in authentication.

5.3 Single Sign-on Protocols

Participants like the speed and convenience of single sign-on, though their qualitative responses also provide details about how existing systems could be improved.

5.3.1 Additional Low-entropy Passwords

Participants liked having a single account that was used to authenticate them to multiple websites. Still, some participants were worried about the risks associated with only having one account for all their websites:

> *"The simplicity is also a downside–after the first log-in, you only have to press 'log in' and it doesn't ask you any verifying information. That doesn't seem like a very secure system. For something inconsequential like a social media site or a blog, I wouldn't mind it, but I want a MUCH more secure authentication system for my bank account. If my google account gets hacked, I assume all the connected accounts that use it to log in can also be jacked. I don't want to take that risk with my important accounts."*

Participants suggest a novel approach to solving this problem. To increase the security of a website, participants propose augmenting single sign-on with a low-entropy password shared with the website (e.g., pin). Security is provided by the high-entropy password of the single sign-on account, yet in the case of an account compromise attackers would be unaware of the low-entropy passwords and be unable to gain access to the website. The cognitive burden for users is also low, as they only need to remember a single high-entropy password, while all other passwords are low-entropy and easily remembered. This is an interesting avenue for future research to explore.

5.3.2 Reputation

With federated single sign-on, the reputation of the provider was key. Qualitative responses from participants indicated that trust in a federated single sign-on system was based on the federating identity provider (IDP) (e.g., Google, Facebook). Participants often cite their opinions of the federating IDP when explaining why they prefer one system to another:

[13]Contrary to normal English usage, transparency in the usable security literature has the opposite meaning, and refers to hiding implementation details from users.

"I would be worried about security. I've heard that Facebook is 'relatively' easy to hack. I would want to be sure that it was all secure before I started using it."

"I trust Google with my passwords."

5.3.3 Dedicated Identity Providers

Some participants prefer that the IDP only handle authentication and not store sensitive information. For example, one participant stated,

"It would be it's own company (not tied to my email, or social network accounts) . . . "

If they were forced to use Google or Facebook as their IDP, one participant indicated that they would create a new account used for authentication only:

"I would make an account separate from my social network and mail specifically for functions like banking etc."

5.4 The Coolness Factor

When participants described what authentication features they were most interested in, they often referred to the "coolness" of that feature. "Coolness" was often related to how different and innovative the technology was perceived to be when compared to current password-based authentication. For example, participants love that Snap2Pass allows them to use their smartphones and obviates the need for passwords:

"Man was that cool!"

"Also, the feel of it made me enjoy doing it. I felt technologically literate and the app felt futuristic as a whole, which I enjoyed."

"I thought the technology was cool. You can snap a code to sign yourself in!"

5.4.1 Biometrics

None of the seven authentication systems we analyzed used biometric-based authentication; nevertheless, over a quarter of participants (28; 29%) discuss biometrics as part of their ideal authentication system. In nearly every case, biometrics were described as being "cool:"

"A fingerprint system would be cool."

"retinal scanner so i just sit in front of my computer and it scans my eye. dope."

Participants liked biometrics because they did not involve an authentication factor that could be forgotten, lost, or stolen:

"The ideal system would scan some part of my body - either eye or thumb - because these are literally ALWAYS with me."

Participants also thought that biometrics were more difficult for hackers to compromise:

"People can hack accounts, but they can't fake your eye-scan pattern"

The list of suggested biometrics is fingerprint, facial, retinal, and voice recognition. While participants may not understand all the implications of biometrics, these results indicate that there is significant interest in adopting biometric-based web authentication. Future research should examine how biometric-based authentication can be implemented on the web while still preserving users' privacy [5].

5.5 Physical Tokens

When using a physical token (i.e., WebTicket, smartphone), participants want to have a fallback mechanism. They are worried that they might lose their phone or WebTicket. They are also concerned with theft, especially when a single token could be used to log in to multiple different accounts or sites. For example, one user stated their concern with Snap2Pass,

"It would make me nervous having all the passwords I need on my phone. For instance, if I forgot or lost it somewhere I could be inconvenienced with having to then make a username and password for all the websites I need, or if it was stolen and the password on my phone compromised somebody could easily access all of my personal and financial information."

Participants also voice concern that if they ever forgot to bring their physical token with them, then they would be unable to log into any websites. Alternatively, some participants also dislike that Snap2Pass requires a smartphone. One participant expresses both concerns in their responses:

"It seems unfortunate that you have to have a smart phone and you also have to have it with you."

5.6 Implementation Lessons

As mentioned in Section 3.2.1, we implemented the seven authentication systems for our studies. We found existing software libraries for Google OAuth 2.0, Facebook Connect, Mozilla Persona, and Snap2Pass that aided our implementation. SAW, Hatchet, and WebTicket were implemented from scratch. The remainder of this section gives lessons learned from implementing the systems.

During authentication, Google OAuth 2.0, Facebook Connect, and SAW use `GET` requests. This caused problems with WordPress, which expects authentication to occur using `POST` requests. We were able to code around this limitation, but this still represents a significant impediment to a clean implementation. It would be best if web authentication proposals allow the use of `POST` requests, as this would reduce development costs.

Google OAuth 2.0 and Facebook Connect both require a security check to prevent impersonation attacks. Facebook Connect's software library handles this check for developers, but Google OAuth 2.0 library requires that developers implement the security check themselves. This check is easy to implement incorrectly, resulting in usability (e.g., failed authentication) and security problems (e.g., impersonation attacks). We recommend that authentication proposals provide publicly available implementations that handle security details for developers.

Implementing WebTicket was straightforward, but the webcam struggled to recognize QR codes. It is unclear if this

problem was a limitation of the webcam or with the current state-of-the-art HTML5 QR code scanning libraries. Regardless, developers need to pay particular attention to this issue if they choose to implement WebTicket or a similar system.

6. RELATED WORK

The Bonneau et al. [6] framework for comparing web authentication schemes includes a set of eight usability benefits that center on convenience and ease-of-use. While a common framework supports a low-barrier method for subjective comparison of alternative system designs, our work emphasizes empirical results from formal user studies.

Dhamija et al. [11] proposed Deja Vu, a graphical password system. The evaluation of Deja Vu included a user study that compared Deja Vu to both passwords and pins. This is the earliest study we identified that compared a proposed authentication system against current password-based authentication. Similar to our work, the study required users in a laboratory to complete assigned tasks. The reported results were both quantitative (completion times and error rates) and qualitative. If a standard metric like SUS had also been included in their study, it would have enabled a direct point of comparison with our study. We want to see this cross comparison become standard practice going forward.

Chiason et al. [10] conducted a 26-person user study comparing two password managers: PwdHash and Password Multiplier. Even though both systems had considered usability issues in some detail in their original papers, the formal study was able to reveal several significant usability challenges. This is another example of why a formal user study must become the norm to complete any system evaluation.

Sun et al. [23] conducted a user study of OpenID, a single sign-on system. The study revealed design flaws in the OpenID login process that led to misconceptions and mistakes due to a lack of transparency of some security details. Similar to our results, some users expressed concern over a single point of failure at the IDP and concerns about the release of private information at the IDP.

To our knowledge, there are only four previous authentication usability studies that utilize SUS. Juang et al. [16] analyzed system-generated mnemonics for remembering passwords and found them more usable than user-generated or no mnemonics. Trewin et al. [26] analyzed three biometric modalities (face, voice, gesture) on mobile phones and compared them to passwords. Passwords were rated the most usable, with gesture and face biometrics slightly lower. Both voice and combinations of biometrics were found to be unusable. Tassabehji and Kamala [25] conducted a user study of a prototype online banking system using biometric authentication, and Bianchi et al. [4] analyzed a system for safely transferring a PIN from a mobile phone in order to authenticate to an ATM. Similar to our research, each of these four studies brought users into a laboratory environment to complete tasks using a prototype system. After using the system, users respond to survey questions about their experience.

Schaub et al. [21] completed a recent study of five graphical passwords systems. The study resulted in a number of helpful insights and guidelines for designers of graphical password systems. Similar to our studies, users were asked to complete a post-study questionnaire after hands-on experience with a system. The questionnaire utilized PSSUQ [17], a standard metric that is not as widely used as SUS.

7. CONCLUSION

Very few proposals for new authentication systems are accompanied by a formal user study. This leaves us with scant empirical data to determine a best-in-class system for the various types of authentication systems or to reason about how the usability of different authentication systems compare against each other. In this paper, we report the results of a series of within-subjects empirical usability studies for seven web authentication systems. Our studies are the first to compare a heterogeneous collection of web authentication proposals.

The result of our studies is that federated single sign-on systems (i.e., Google OAuth 2.0, Facebook Connect, Mozilla Persona) and Snap2Pass are rated as having the best overall usability. Our results validate SUS as an appropriate metric for comparing the usability of authentication systems, namely because the SUS score for a given system was consistent across different participant groups and proved to be a strong indicator of users' preferences.

Our usability studies also gathered insightful information from participants' qualitative responses. We found that transparent authentication systems are rated as usable but also lead to confusion and a lack of trust from users. Additionally, while participants rate the usability of single sign-on highly, they are interested in augmenting it with additional site-specific, low-entropy passwords. Our results show that over half of participants are willing to use new authentication systems in their everyday life, but that they are most interested in adopting systems that they perceive as different and innovative (e.g., biometrics, Snap2Pass).

Finally, our results have significant implications moving forward. First, the security and usability communities should collaborate more effectively on authentication proposals to ensure that new systems are rigorously evaluated in terms of both security and usability. Second, usability studies that incorporate SUS should become a standard practice for vetting all new authentication proposals. New authentication systems should meet a minimum SUS score before receiving serious consideration. Based on our experience and the literature on SUS, a minimum score of 70 is a reasonable expectation. Widespread adoption of these practices would lead to a significant leap forward for authentication.

8. ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers, Jay McCarthy, Charles Knutson, and Daniel Zappala for feedback on earlier drafts of the paper.

9. REFERENCES

[1] Steam Guard. `https://support.steampowered.com/kb_article.php?ref=4020-ALZM-5519`. [Online; accessed 2014/11/20].

[2] A. Bangor, P. Kortum, and J. Miller. An empirical evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, 24(6):574–594, 2008.

[3] A. Bangor, P. Kortum, and J. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[4] A. Bianchi, I. Oakley, and D. S. Kwon. Using mobile device screens for authentication. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, OzCHI '11, pages 50–53, New York, NY, USA, 2011. ACM.

[5] J. Bonneau, E. W. Felten, P. Mittal, and A. Narayanan. Privacy concerns of implicit secondary factors for web authentication. In *SOUPS Workshop on "Who are you?!": Adventures in Authentication*, 2014.

[6] 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*, pages 553–567. IEEE, 2012.

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

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

[9] S. Chiasson, E. Stobert, A. Forget, R. Biddle, and P. C. Van Oorschot. Persuasive cued click-points: Design, implementation, and evaluation of a knowledge-based authentication mechanism. *IEEE Transactions on Dependable and Secure Computing*, 9(2):222–235, 2012.

[10] S. Chiasson, P. C. van Oorschot, and R. Biddle. A usability study and critique of two password managers. In *USENIX Security*, 2006.

[11] R. Dhamija and A. Perrig. Deja Vu — a user study: Using images for authentication. In *USENIX Security*, 2000.

[12] 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*, pages 17–38. Springer, 2012.

[13] S. L. Garfinkel. Email-based identification and authentication: An alternative to PKI? In *Symposium on Security and Privacy*, pages 20–26. IEEE, 2003.

[14] E. Hayashi, B. Pendleton, F. Ozenc, and J. Hong. WebTicket: Account management using printable tokens. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 997–1006. ACM, 2012.

[15] R. Jhawar, P. Inglesant, N. Courtois, and M. A. Sasse. Make mine a quadruple: Strengthening the security of graphical one-time PIN authentication. In *International Conference on Network and System Security*, pages 81–88. IEEE, 2011.

[16] K. A. Juang, S. Ranganayakulu, and J. S. Greenstein. Using system-generated mnemonics to improve the usability and security of password authentication. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 56, pages 506–510. SAGE Publications, 2012.

[17] J. R. Lewis. Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.

[18] C. Robison, S. Ruoti, T. W. van der Horst, and K. E. Seamons. Private facebook chat. In *International Conference on Privacy, Security, Risk and Trust and International Confernece on Social Computing*, pages 451–460. IEEE, 2012.

[19] S. Ruoti. Authentication melee: A usability analysis of seven web authentication systems. Master's thesis, Brigham Young University, 2015. `http://scholarsarchive.byu.edu/etd/4376/`.

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

[21] F. Schaub, M. Walch, B. Könings, and M. Weber. Exploring the design space of graphical passwords on smartphones. In *Symposium on Usable Privacy and Security*. ACM, 2013.

[22] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The Emperor's new security indicators. In *Security and Privacy*, pages 51–65. IEEE, 2007.

[23] S.-T. Sun, E. Pospisil, I. Muslukhov, N. Dindar, K. Hawkey, and K. Beznosov. What makes users refuse web single sign-on?: An empirical investigation of OpenID. In *Symposium on Usable Privacy and Security*. ACM, 2011.

[24] H. Tao. *Pass-Go, A New Graphical Password Scheme*. PhD thesis, University of Ottawa, 2006.

[25] R. Tassabehji and M. A. Kamala. Evaluating biometrics for online banking: The case for usability. *International Journal of Information Management*, 32(5):489–494, 2012.

[26] S. Trewin, C. Swart, L. Koved, J. Martino, K. Singh, and S. Ben-David. Biometric authentication on a mobile device: a study of user effort, error and task disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 159–168. ACM, 2012.

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

[28] T. W. van der Horst and K. E. Seamons. Simple authentication for the web. In *International Conference on Security and Privacy in Communications Networks and the Workshops*, pages 473–482. IEEE, 2007.

[29] D. Weinshall. Cognitive authentication schemes safe against spyware. In *Symposium on Security and Privacy*, pages 295–300. IEEE, 2006.