A Dual-Task Interference Game-Based
Experimental Framework for Comparing the
Usability of Authentication Methods

Jean-Baptiste Subils\textsuperscript{1}, Joseph Perez\textsuperscript{2}, Peiwei Liu\textsuperscript{2}, Shamaria Engram\textsuperscript{1}, Cagri Cetin\textsuperscript{1}, Dmitry Goldgof\textsuperscript{1}, Natalie Ebner\textsuperscript{2}, Daniela Oliveira\textsuperscript{3}, Jay Ligatti\textsuperscript{1}

\textsuperscript{1} Department of Computer Science and Engineering, University of South Florida
\textsuperscript{2} Department of Psychology, University of Florida
\textsuperscript{3} Department of Electrical and Computer Engineering, University of Florida

Abstract—This paper introduces a game-based framework to compare the usability of authentication methods. The framework uses a dual-task interference technique to determine the usability of authentication methods. In the experiment, subjects participate in a multi-tasking game that simulates a conversation being interrupted by authentication requirements. By simulating a conversation, the goal is to reproduce a real use of authentication, and collect ecologically sound data. Participants also perform each authentication method in a standalone manner, which allows for comparison of the usability under two different cognitive loads. The authentication techniques evaluated represent each of the three main authentication factors, specifically password, fingerprint, and coauthentication. The three aspects of usability used to compare authentication techniques in this framework are efficiency, effectiveness, and satisfaction. An experiment with 43 participants enrolled was conducted to collect data pertaining to these aspects. The results show that fingerprint and coauthentication (both laptop and phone) are the more usable techniques evaluated.

Index Terms—security, authentication, usability

I. INTRODUCTION

Authentication is one of the most common security activities end-users perform. Due to this activity being reoccurring, time consuming, and often considered annoying [9], users tend to prefer authentication with minimal effort on their part. Thus, in order for end-users to adopt a new authentication method, usability is crucial. Due to the usability and popularity of passwords, a novel authentication method also needs to yield significantly better results than a password-based authentication method [2].

Authentication methods are based on three standard factors. For example, the inherence factor can be facial recognition, the knowledge factor can be a password, and the possession factor can be a One Time Password (OTP) sent to the user’s phone. Due to the fundamental differences between these factors, comparing authentication methods using different factor(s), especially in terms of usability, can be complicated.

Few tools are available to compare authentication methods. Bonneau et al. introduced a subjective framework which provides a qualitative scale to compare authentication methods [2]. While offering a good assessment of authentication methods under a thorough list of categories, this framework remains a high-level overview for each category (e.g., usability). Usability is highly subjective, and thus, depends on the perception of users. Therefore, feedback from participants is necessary to obtain a more precise comparison. Another standard approach to compare authentication methods is the System Usability Scale (SUS), which is an effective metric for usability comparison [3], [15]. However, SUS only captures the satisfaction aspect of usability. SUS collects feedback from users through a questionnaire answered via a Likert scale that provides a score from $0 \rightarrow 100$ [1].

This paper aims to provide an experimental framework as a solution to compare all authentication methods in terms of usability, regardless of the authentication factor(s). The framework incorporates the System Usability Scale and other metrics (e.g., time of completion, accuracy) to compare authentication methods.

A. The Framework’s Motivation & Background

This section introduces the design principles of the framework. The framework presented in this paper focuses on the usability aspect of authentication methods and incorporates multiple metrics that can be applied for usability comparison.

To collect ecologically sound data, a scenario close to reality needs to be designed. Users do not choose to authenticate but are, instead, interrupted by authentication requirement(s). Authentication happens when users are accessing some resources requiring them to prove their identity. Thus, an activity representing a specific action being interfered with should be defined to reproduce a real use of an authentication method. Participants can attempt to complete the activity while being obstructed by authentication requirements. For example, users may be required to authenticate when accessing a website, while calling a support service, participating in a teleconference, or carrying on a conversation.

Authentication usually interferes with the access of some resources, thus, dual-task interference is a suitable technique to study user perception of authentication methods [12]. In the field of cognitive psychology, dual-task interference is used to determine a person’s cognitive load and ability to multitask. Due to the results of dual-task interference, the framework
requires participants to undergo a dual-task interference game in addition to performing each authentication technique in a standalone manner.

The framework should collect information on authentication methods performed in a standalone manner to provide a baseline to compare with the data collected during the dual-task interference game (i.e., DTI game). This part serves as a training phase and allows participants to get accustomed to each of the authentication techniques evaluated. Prior to the DTI game, another training phase is helpful to familiarize participants with the game. The data resulting from this training can also be helpful in detecting improvements. The training phases should require participants to complete, first the activity, second the authentication tasks, and finally, both simultaneously to engage with the DTI game.

Gamification is a compelling incentive that pushes participants to complete experiments, and, in turn, produces meaningful data in case studies [6], [7]. The experiment includes a multi-tasking game to actively engage participants. A scoreboard is a type of gamification element that gives feedback to the participant on their performance and provides an incentive to surpass themselves. The participant’s score should be updated in real time to provide direct feedback. Participants’ compensation can be based on their score, as an additional compelling incentive. Figure 1 shows a screenshot of the web application used, with the progress bar on the top left, the participant’s score in the circle on the top right, and in the middle the authentication requirement.

B. The Framework’s Contributions

The framework is used to evaluate the following usability aspects of authentication methods:

- **Efficiency**, which is the authentication’s completion time.
- **Effectiveness**, which is the success rate.
- **Satisfaction**, which is rated via participants’ feedback.

Efficiency is defined as the length of time necessary for a user to be authenticated. Thus, this time is calculated from the first user action performed to authenticate until the user receives the authentication result. Additionally, this time takes into account the user-interaction time required to perform an authentication task.

To compare authentication methods fairly, regardless of the authentication factor and in terms of usability, a success rate is essential. False positives directly affect the usability of an authentication method, and in practice, re-authenticating due to a failed attempt, with a certain limit to prevent brute-force attacks, is allowed to improve usability. The purpose for allowing retries is to reduce the inconvenience induced from forcing users to re-authenticate. False negative and false positive rates are metrics used to determine the security and usability of biometric authentication methods [4]. However, other types of authentication factors are less subject to accuracy problems. Accuracy will be used as a metric to determine the effectiveness of each authentication method evaluated.

Satisfaction is the subjective perception of usability. To collect participants’ feedback the framework uses the Authentication Experience Questionnaire, which includes the System Usability Scale (SUS) questionnaire and various feedback questions. SUS is a widely used and accepted approach to determine the usability of a computer system [3].

II. RELATED WORK

This section elaborates on the concepts of Dual-Task Interference and Gamification and discusses the related work pertinent to our study and related to authentication usability.

A. Dual-Task Interference

Many research studies have concluded that Dual-Task Interference affects response time and general task performance [10], [13]. Humans traditionally struggle when faced with two discrete, but simultaneous tasks that require independent responses. The resulting delayed response has been attributed to many different psychological and cognitive phenomena. Most explanations refer to the psychological refractory period (PRP), a bottleneck effect in later cognitive processes that prevents a second stimulus from being processed until an initial stimulus is processed [13]. While the source of this cognitive bottleneck is contested, its ability to create strain and to delay responses in human cognition is widely accepted [21]. Research has also demonstrated that this delayed response effect can be minimized under appropriate training and adaptive executive control [17].

The purpose of the framework presented in this paper is to more accurately simulate the daily experience of authenticating while simultaneously performing an interactive task. Authenticating while performing another task presents a larger cognitive load than authenticating alone and, therefore, needs to be appropriately accounted for.

B. Gamification

Video games attract players through immersive and unique mechanics that also encourage players to remain actively engaged; gamification is the process of applying these mechanics to other contexts to provide similar benefits [5]. Gamification incorporates concepts such as point scoring, leaderboards, badges, and achievements for completing certain tasks to enhance a player’s immersive experience. Incorporating such
concepts gained popularity due to the potential benefits, including actively engaging users [6], [7]. Through active engagement fostered by gamification, experimental studies can produce more meaningful data [6].

The gaming principle used in the study presented in this paper is a scoreboard based on the successful completion of both tasks (authentication and user activity). Updating the scoreboard allows for real-time feedback, which then also encourages participants to improve their score. As an additional incentive, participants’ compensation is based on their performance (i.e., score).

C. Methodology of Evaluating Authentication Usability

A large body of literature pertaining to usability of computer systems is available; however, comparing the usability of authentication methods remains difficult due to the different techniques available. The lack of a standard method evaluating authentication usability, and some of the different methods proposed, will next be discussed.

Comparing authentication methods is often achieved by comparing methods as a whole, which results in a high-level overview of usability [2]. The Quest to Replace Passwords [2] provides a qualitative scale to compare authentication and encapsulates the three main principles of usability: efficiency, effectiveness, and satisfaction. However, this scale is difficult to assess objectively to compare authentication methods, due to the subjective nature of the criteria [2, Section V-B].

The three main aspects of usability (efficiency, effectiveness, and satisfaction) should be considered to properly assess a system’s usability. Indeed, these aspects are not always correlated, and assumptions on the overall system’s usability may not be accurate [8].

The satisfaction aspect of an authentication method is often the main research focus as the adoption of an authentication method depends on end-users. The System Usability Scale (SUS) is considered a standard to collect this satisfaction measurement via a study collecting participants’ perceptions [3], [9], [15]. The questions from the SUS questionnaire are answered via a Likert scale that allows for the calculation of a usability score from 0 to 100 [1]. This score has been assessed for a wide variety of computer systems and revealed to be consistent and reliable [14], [20]. SUS is also a recommended metric standard to compare authentication methods [15].

As is well known, passwords are the predominant authentication method, and an extensive literature has evaluated their usability over the last few decades. However, most usability research specific to passwords is not directly applicable to other authentication methods. The research mostly pertains to improving the usability of passwords. Recent studies assessed on passwords often focus on the effects of password policies on usability [18]. Other recent studies also focus on typing speed due to the emergence of smartphones and the need to enter passwords on virtual keyboards [16], [22]. In these studies the experimental setup is similar where participants have to type passwords and answer questionnaires to collect feedback and demographic data. Depending on the goal of the study, participants are given the password(s) or are required to create a password(s).

The focus of biometric usability research is similar to the research on password usability in that a biometric method’s usability is typically only compared to other biometric authentication methods. Biometric methods are comparable via a False Acceptance Rate (FAR) and a False Rejection Rate (FRR) [4], [19]. The FAR is often used to indicate a level of security while the FRR is often used to indicate a level of usability. These clearly defined metrics are the reason why biometric methods are easily comparable. However, FRR rates only indicate the effectiveness of the authentication methods. Because effectiveness and satisfaction are not correlated [8, Section 4.5], effectiveness is insufficient for determining a usability score that can be compared to other authentication methods (i.e., using other authentication factor(s)).

III. METHODOLOGY

This IRB-approved study was conducted in a lab where each participant followed instructions from a web application on a laptop. A researcher also guided them through the procedure and answered questions. This section will present the authentication techniques evaluated; then discuss the recruitment process and the resulting demographics; then we will detail what participants underwent, the specific hardware used, and finally the limitations associated with such a study.

A. Authentication Techniques Evaluated

The authentication techniques evaluated in this paper’s framework represent the three main authentication factors (i.e., knowledge, inheritance, possession). Passwords were chosen to represent the knowledge factor because of their popularity. A fingerprint authentication method was chosen to represent the inheritance factor, because of the availability of fingerprint sensors on mobile devices, and because fingerprint authentication has been well researched. Coauthentication was chosen to represent the possession factor because this authentication method’s usability has never been evaluated [11].

The coauthentication and password authentication methods were completed on two different input devices: a laptop and a phone. Therefore, including fingerprint authentication on a phone, a total of five authentication techniques were evaluated. Testing on two devices allowed for comparison between input devices.

Each of these authentication methods requires a user registration prior to authenticating, in order to register a password, fingerprint, or device keys (cryptographic secrets used for coauthentication). Thus, the registration phase was completed prior to the start of the experiment.

B. Study Recruitment and Demographics

Participants for this study were recruited in various ways. Primarily, the cloud-based participant pool management software SONA Systems was used through the University of Florida’s Psychology department to recruit students enrolled in the general psychology course. Students taking the course
TABLE I: Demographics of the 43 participants enrolled in the study.

<table>
<thead>
<tr>
<th>Demographic Category</th>
<th># Participants (N = 43)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>10</td>
<td>23%</td>
</tr>
<tr>
<td>Female</td>
<td>33</td>
<td>77%</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 years old</td>
<td>14</td>
<td>32%</td>
</tr>
<tr>
<td>19 years old</td>
<td>16</td>
<td>37%</td>
</tr>
<tr>
<td>20 years old</td>
<td>9</td>
<td>21%</td>
</tr>
<tr>
<td>21 years old</td>
<td>4</td>
<td>9%</td>
</tr>
<tr>
<td>Ethnicity and Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>9</td>
<td>21%</td>
</tr>
<tr>
<td>Black</td>
<td>4</td>
<td>9%</td>
</tr>
<tr>
<td>Asian</td>
<td>18</td>
<td>42%</td>
</tr>
<tr>
<td>White</td>
<td>28</td>
<td>64%</td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual</td>
<td>20</td>
<td>47%</td>
</tr>
<tr>
<td>Native English</td>
<td>36</td>
<td>83%</td>
</tr>
</tbody>
</table>

were required to sign up for studies, and received 4 credits for participation in addition to the extra compensation based on task performance. Additional participants were recruited using flyers.

Table I shows the demographics of the 43 participants enrolled in the study. All interested participants were accepted; however, the recruitment methods attracted primarily college students (all under 22 years old). Additionally, 36 out of the 43 total participants were female, so this study has a disproportionate representation of the female demographic.

C. Study Design

According to the framework guidelines, described previously in Section I-A, the participants completed the following steps:

(a) The Participant Information Questionnaire
(b) A training phase for the authentication techniques (i.e., standalone)
(c) A training phase for the user activity
(d) The Authentication Experience Questionnaire to collect data on authentication alone
(e) A training phase for the Dual-Task Interference game (i.e., DTI game)
(f) The DTI game (Administered in six sessions)
(g) The Authentication Experience Questionnaire a second time to collect data on authenticating while multi-tasking

For this study we chose to make the user activity simulate a conversation, to represent the common use case of authentication interrupting conversation. To mimic a conversation, participants repeated a series of words. The accuracy of correctly repeating these words was recorded, and participants acquired two points for each correct word. Each conversation lasted five minutes and used the same series of words. There were a total of six conversations, for a total of thirty minutes. For the remainder of this paper, these conversations will be described as the DTI game or multi-tasking game.

Participants also accumulated more points by successfully performing the authentications required. Each successful authentication earned ten points. In order for the participants to know each authentication result, it was displayed for two seconds.

To get familiarized with the various components of the experiment, participants went through three training phases. In the first training phase, participants repeated a series of words for thirty seconds to simulate a conversation. The second phase required the participants to perform each of the authentication methods twice. In the third training phase participants had to repeat words while authenticating for one and a half minutes (i.e., practice the multi-tasking game).

To collect participants’ feedback, the Authentication Experience Questionnaire (AEQ), was given twice. The questionnaire was given once after participants performed the first training phase (i.e., authentication methods training), and a second time after the multi-tasking game. This repetition allows for comparison between authentication performed in a standalone manner versus during the DTI game.

The participant’s score was the only incentive for participants to perform the experiment properly. The score was updated and displayed in real time during the game. After each conversation of the game the participant could see the score earned and take a break. At the end of the experiment the compensation was calculated from the best score obtained between all six games. Each successful authentication earned 10 points and each successful audio task earned 2 points.

D. Hardware Specification

To complete the experiment, participants used a laptop and a smart-phone provided. The server (i.e., authenticator) was also deployed on the same laptop. The following is the hardware specification of these devices:

- Laptop: Dell Windows 10, memory 8GB, processor i5-7200U 2 cores at 2.5 GHz, and a 13 inch screen size.
- Phone: LG V20 with 4GB of memory, a 1.6GHz quad-core processor, Android 7, and a 5.7 inch screen size.

One relevant specification of the hardware here is the placement of the fingerprint scanner, which was located on the back of the smart-phone.

E. Limitations

The demographic data shows that most participants were female (73%), college students (100%), and relatively young (100% are under 22 years old). Thus, the data obtained is more useful at predicting usability in female college students than any other group.

Technologically, each individual’s comfort with the specific hardware aspects of the experiment can be considered a confounding variable. Many of our participants use Apple products such as the iPhone and Mac computer. The level of comfort these participants had with our Android phone and Windows computer may not match that which they typically feel for their personal devices. For example, the screen size of both the laptop and the phone may differ from the ones the participants are used to. Additionally the collection of each participant’s response to an authentication task was
TABLE II: Comparison of completion times for each authentication technique evaluated.

<table>
<thead>
<tr>
<th>Authentication method</th>
<th>Completion times (seconds)</th>
<th>Authentication Success Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standalone Average</td>
<td>Median</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>3.25</td>
<td>2.28</td>
</tr>
<tr>
<td>Password (laptop)</td>
<td>8.83</td>
<td>8.12</td>
</tr>
<tr>
<td>Password (phone)</td>
<td>9.25</td>
<td>8.75</td>
</tr>
<tr>
<td>Coauthentication (laptop)</td>
<td>0.68</td>
<td>0.63</td>
</tr>
<tr>
<td>Coauthentication (phone)</td>
<td>1.09</td>
<td>0.92</td>
</tr>
</tbody>
</table>

TABLE III: Success rates for fingerprint and password authentication techniques weighted per participant.

<table>
<thead>
<tr>
<th>Authentication method</th>
<th>Standalone Success Rate (%)</th>
<th>Multi-tasking Success Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fingerprint</td>
<td>99.95</td>
<td>99.99</td>
</tr>
<tr>
<td>Password (laptop)</td>
<td>99.92</td>
<td>99.95</td>
</tr>
<tr>
<td>Password (phone)</td>
<td>99.86</td>
<td>99.95</td>
</tr>
</tbody>
</table>

TABLE IV: Averages of System Usability Scale scores of the authentication methods evaluated.

<table>
<thead>
<tr>
<th>Authentication method</th>
<th>SUS Score Standalone</th>
<th>Multi-tasking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fingerprint</td>
<td>88</td>
<td>82</td>
</tr>
<tr>
<td>Password (laptop)</td>
<td>81</td>
<td>76</td>
</tr>
<tr>
<td>Password (phone)</td>
<td>78</td>
<td>74</td>
</tr>
<tr>
<td>Coauthentication (laptop)</td>
<td>81</td>
<td>82</td>
</tr>
<tr>
<td>Coauthentication (phone)</td>
<td>81</td>
<td>82</td>
</tr>
</tbody>
</table>

The results show that fingerprint and coauthentication (both laptop and phone) are the more usable techniques evaluated. Their satisfaction and efficiency results are significantly better.
than passwords, though we are unable to draw conclusions regarding the effectiveness due to the similarity in results. Coauthentication yielded higher efficiency results than fingerprint. However, the satisfaction results of fingerprint are overall better or as good as coauthentication’s (both laptop and phone).

The framework enables authentication methods’ usability, including efficiency, effectiveness, and satisfaction, to be evaluated uniformly using a standard methodology even across varying authentication factors.

Several extensions exist for using the framework to evaluate authentication methods’ usability in different contexts, by varying the user activities, game design principles, or the authentication methods themselves.

Many user activities are obstructed by authentication requirements everyday, thus ensuing studies could modify the difficulty and the type of the activity simulated.

The activity simulated in the study presented was a conversation; however, simulating a conversation by having participants repeat words may not require enough cognitive load. The difficulty of this task can be adjusted to collect data and investigate the resulting effect(s). The following are potential modifications that would result in a different difficulty level:

(a) The sets of words can be categorized based on a difficulty level.

(b) The speed of the audio can be accelerated to increase the difficulty.

(c) The sets of words during the experiment can appear in a randomized order.

(d) The auditory task can be replaced with full sentences or questions.

All these combinations are avenues to explore, to determine the effect of Dual-Task Interference and cognitive load while authenticating.

Additional game-design concepts can be included into the framework to further engage participant interest. For example, a leaderboard displayed during each break can give feedback to participants on their performance compared to each other. Several participants, during the trials, expressed interest in knowing how their scores compared to previous subjects. This particular gamification design is, therefore, one that should be considered in future related studies.

Another possible extension relates to setting a password for participants. In the experiment presented here, the participants’ password was given, which was meant to prevent any “weak” password creation [22]. However, it cannot yet be determined whether having participants create their own passwords would significantly impact the study’s results.

VI. ACKNOWLEDGEMENT

This research was funded by a Collaborative Seed Award from the Florida Center for Cybersecurity.

REFERENCES


