5-2012

A SYSTEM-GENERATED PASSWORD AND MNEMONIC APPROACH TO OPTIMIZE THE SECURITY AND USABILITY OF TEXT-BASED PASSWORDS

Sanjaykumar Ranganayakulu
Clemson University, sanjay6788@gmail.com

A SYSTEM-GENERATED PASSWORD AND MNEMONIC APPROACH TO OPTIMIZE THE SECURITY AND USABILITY OF TEXT-BASED PASSWORDS

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Industrial Engineering

by
Sanjaykumar Ranganayakulu
May 2012

Accepted by:
Dr. Joel S. Greenstein, Committee Chair
Dr. Anand K. Gramopadhye
Dr. Byung Rae Cho

ABSTRACT

In this study a novel password generation policy called the system-generated password and mnemonic was designed and implemented. The intent of this policy was to optimize both the security and usability of text-based passwords. After implementing the policy we evaluated its usability and compared it with three other existing policies: user-generated password, system-generated password and user-generated mnemonic for a system-generated password. In order to have a fair comparison among the policies we maintained a constant level of security of 30±2 entropy as dictated by NIST level 2 standards.

The study involved 64 participants, equally divided into four groups, 16 in each password policy condition. The study took place over two sessions, with a period of 5-7 days in between them. In the first session, depending on the password policy condition, the participants were either assigned or asked to create a password. The participants were then asked to recall their passwords in the same session and after 5-7 days in the second session. The four password policy conditions were compared with respect to the following dependent variables: the time taken to create the password account, the password creation error rate, the time taken to recall and recall error rates for both sessions, unrecoverable passwords in the second session, proximity of the recalled password to the stored password as measured by the Damerau-Levenshtein and Jaro-Winkler edit distances; and the subjective ratings for the NASA task load indices and the System Usability Scale questionnaire.

There was a significant effect of password policy condition on the time taken to create a password account and for the performance index of the NASA-TLX questionnaire. Across the task sessions, there were statistically significant differences for the time taken to recall the password, recall error rates, the performance index of the NASA-TLX questionnaire and the SUS score. There were no significant differences for creation error rates, creation SUS, recall error rates and unrecoverable passwords among the password policy conditions.

The results of this study suggest that overall performance was better for the user-generated policies (user-generated password and system-generated password along with a user-generated mnemonic) than for the system-generated policies (system-generated password and system-generated password and mnemonic). One of the reasons for this result might be that the direct involvement of the user in generating the password or mnemonic enhances their memorability. Other reasons mentioned by the users were that the system-generated mnemonic policy was complex and employed difficult words which were difficult to memorize and thus recollect. As a result of conducting this experiment it is concluded that user-generated policies are better in terms of usability and memorability than system-generated passwords. However, the user feedback recorded in this study suggests a number of approaches for improving the usability of system-generated password policies.

DEDICATION

The thesis is dedicated to my beloved parents, Dr. Ranganayakulu and Mrs. Hemalatha Ranganayakulu; my brothers G. R. Diveshwar, R. Lokeshwaran; my grand-parents R. Mohan Gopal and Yamuna Mohan; my uncle D. Ravi, aunt R. Uma Devi and God Almighty.

ACKNOWLEDGEMENTS

I would like to express my heartfelt thanks for the constant support extended by my advisor Dr. Joel S. Greenstein. This work would not have been possible but for him. I also am deeply indebted to my committee members Dr. Anand K. Gramopadhye and Dr. Byung Rae Cho for their technical support.

I am also thankful to Mrs. Barbara Ramirez of the English department for helping me to organize my flow of thoughts into the right words.

I would also like to express my gratitude to my lab mates, Kevin Juang, Kapil Chalil Madathil, Sourav Bhuyan, Meera Ramachandran and Sumonthip Chompoodang. Kevin Juang helped me develop my application and Meera Ramachandran helped me recruit participants for my study and editing my written thesis. I am deeply indebted to them for their encouragement and their continued help with all their comments right from the start of my thesis to the end of it.

I am thankful to all my friends and roommates at Clemson, especially Vineeth Kumar Jampala and Nikhil Kumar Seera, who have helped me through this journey making it a memorable experience.

Lastly but most importantly, my parents are responsible for their never ending emotional support and encouragement always when I needed them. I am also indebted to my all other family members and teachers who helped me throughout the phases of my education.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

In 1961 when MIT developed the Compatible Time-Sharing System (CTSS), passwords were first used in computers to authenticate the users. Since then their increased use for personal purposes has led to privacy issues being taken increasingly seriously. This advent of personal computers and the introduction of the World Wide Web (WWW) have resulted in a proliferation of personalized web application services. As these applications contain the private information of users, they are protected by authentication mechanisms to constrain access to only legitimate users. Brostoff and Sasse (2000) classified the authentication processes to identify users broadly into three types:

1. **Knowledge-based authentication** uses a secret word or phrase shared between the user and the computer system, e.g. text-based passwords.
2. **Token-based authentication** uses a physical token that is difficult to obtain or forge, e.g. ATM cards or ID cards with magnetic strips.
3. **Biometric authentication** relies on unique details of a person's anatomy or behavior, matching the electronic equivalent of those characteristics to the users, e.g. retinal scan, finger print reader, voice recognition.

Currently, knowledge-based authentication mechanisms like text-based passwords are used more widely than the others because they were the first developed and they do not require special equipment; they will probably continue to be so for the foreseeable future primarily because of user resistance to change and the cost of modifying existing systems. Therefore, this research study focuses on ways to improve security and usability of the text-based password.

The security of any authentication system is directly proportional to the difficulty with which an adversary can obtain illegal access into the system (Jeyaraman & Topkara, 2005). For example, text-based password that is difficult to crack could be intuitively thought of as a string that is not based on a dictionary word and has maximum entropy ("looks" totally random) (Morris & Thompson, 1979). However, the ability to remember a completely unrelated sequence of items is very limited. Hence, the more secure the password is (the greater its randomness), the more difficult it is for users to remember. This limited ability is further taxed by the fact that a typical user has access to multiple computer system applications and is advised to use a unique password for each. Secure website account providers like banks and universities impose restrictions on their users' log-in passwords. These restrictions are not standardized; for example, some websites ask users to incorporate at least one special character and a number in a password of a specified minimum length, and others ask the users to have at least an uppercase letter and at least one number in the password. This practice, although it enhances security, adversely affects website usability because users have difficulty remembering a variety of passwords constructed to satisfy different requirements. These issues suggest that text-based password authentication systems require further improvement to make them usable while maintaining high levels of security.

In general text-based passwords can be classified into two types: user-generated and system-generated. User-generated passwords have been found to be less secure but more easily remembered than system-generated ones because they are often words or phrases having personal meaning (Proctor, Mei-ching Lien, Vu, Schultz, & Salvendy, 2002). On the other hand, system-generated passwords are considered to be more secure but less easy to remember because they tend to be random. To address this issue, researchers (Klein, 1990) have proposed a third policy in which users generate a seemingly random password from a mnemonic phrase, which then serves as a memory aid. However, Kuo et al. (Kuo, Romanosky, & Cranor, 2006) found that these user-generated mnemonic-based passwords are not as secure as randomly generated ones because users tend to choose popular phrases found easily on the Internet. For their study, they created a relatively small database of such popular phrases and found that it could crack 5% of the passwords created by the participants in the study. Even though this percentage is small, the researchers suggest that a larger database would increase the probability that this type of password could be cracked. The study proposed here investigates the use of software to generate random passwords along with a mnemonic aid for the users to help them easily remember their passwords. This password generation policy is compared with other password generation policies: user-generated passwords with restrictions, system-generated random passwords with no mnemonic assistance and system-generated random passwords with mnemonic training provided to the users.

Specifically, this study evaluates these four types of password generation policies in terms of usability while maintaining a security standard dictated by NIST level 2 guidelines (Burr, Dodson, & Polk, April 2006). The metrics used to measure the usability of the policies are:

- **Password retention accuracy** - measures the accuracy with which participants recall their password by calculating the Damerau–Levenshtein edit distance and Jaro-Winkler proximity edit distance of the recalled password in comparison with the correct password.
- **Password creation and memorization time -** measures the time taken by participants to create and/or memorize their passwords.
- **Password creation/recall error rate** - is the ratio between the total number of unsuccessful password creation/recall attempts and the total number of attempts made by the participants to successfully create/recall their password. If the users cannot recall their passwords after a specified number of attempts, then the error rate is recorded as 1. If the user successfully recalls the password in his/her first attempt, then the error rate is recorded as 0.
- **Workload index measure -** is the demand perceived by the users while creating a password in the first session and while recalling and using it in the second session.
- **Subjective satisfaction measure -** is recorded using the System Usability Scale (SUS) questionnaire which indicates the level of user satisfaction with the password generation policy.

2. LITERATURE REVIEW

Most research in the area of usability in computer security has compared different policies of password generation by investigating usability and security separately or in combination. In early research in this area, Zviran and Haga (1993) conducted a usability study comparing user-generated and randomly generated passwords. Using questionnaires they asked 106 participants to generate and record passwords. Then, the participants were also given a randomly generated password to memorize. This within subject design found that after a three-month interval, 35% recollected their self-generated passwords correctly, but only 23% recalled their assigned random passwords.

Similar to Zviran and Haga's work, Bunnell, Podd, Henderson, Napier, & Kennedy-Moffat (1997) compared the retention and guessing rate of user-generated and assigned passwords. This study was based on a questionnaire designed for two sets of participants. The first set, the main respondents, was directly contacted by the researchers. The second set of participants, referred to as significant others, was chosen by the main respondents. The main respondents were tested to determine the retention rate of self-generated and assigned passwords, while the significant others were tested to determine the guessability of the passwords generated by and assigned to the first set. In addition to demographic information, the questionnaire provided to the main respondents collected answers to 20 fact-based and 20 opinion-based questions. It concluded by asking the participants to generate new passwords without any restrictions and assigning each a second experimenter-generated password. These assigned passwords, which were not completely random, consisted of 8 characters, a three-letter word followed by a numeral from 1 to 9 and then a four-letter word, e.g. end5aide or fit4make. After a two-week interval, the main respondents were given a second questionnaire, asking them to recollect both passwords. The self-generated passwords were recalled correctly by 77% of the main respondents and the assigned passwords by 70% of them. These results suggest that the former were somewhat more easily recalled than the latter even though the assigned passwords were designed to be easy-to-remember and were not random nonsense words.

To determine the guessability rate of these passwords, a separate questionnaire was used for the significant others, requiring them to guess the answers given to the questions asked of their respective main respondents. They were also asked to guess both passwords. Overall, 5% of the significant others correctly guessed the self-generated password, but none guessed the assigned password. These results suggest that assigned passwords are more secure against brute force and social engineering attacks than self-generated ones. However, the self-generated passwords did not have any restrictions, so the users may have generated less secure ones easily guessed by others.

Extending Bunnell et al.'s work, Pond, Podd, Bunnell, & Henderson (2000) focused on testing the recall and guessing rates for a word association password generation technique where the user is given or chooses a word to use as a cue for generating a second word. The response to the cue word acts as a password. Using a methodology similar to Bunnell et al., they determined the recall and guessing rates of three such word association password generation policies: response only, cue and response, and theme. In the response only group, respondents were required to generate an associated response for each of 20 cues. In the cue and response group, respondents generated both cues and associated responses, while in the theme group respondents generated both cue and response words having first decided upon a theme for their word associations. This between subject study did not show any significant differences in recall and guessing rates among the three policies tested. Sixty-nine percent of the participants in the response only group, 61% of the cue and response group and 73% of the themes group recalled their passwords correctly.

Keith, Shao, & Steinbart (2007) compared user-generated password policies with minimal restrictions, high restrictions and passphrases. In general the passphrase consists of a group of words which acts as a password instead of a group of characters as in the case of typical passwords. This study, which employed a more realistic password use environment than Pond (2000), measured log-in success and typographical error rates. This between subject design was conducted over a period of 12 weeks, with participants logging in regularly to access the author-created web application. The overall log-in success rates were highest for the user-generated minimal restriction policy at 85.61%, followed by the user-generated high restriction policy at 80.38% and passphrases at 71.58%. These results were supported by a participant satisfaction survey, ranking user-generated minimal restriction first, followed by user-generated high restriction and passphrase passwords.

Leonhard & Venkatakrishnan (2007) compared three random password generators, ALPHANUM, DICEWARE, and PRONOUNCE3. This between subject study required the participants to complete a questionnaire that included a screen shot of a fictional website before assigning each of them a password randomly generated by one of the three policies. After two weeks a second questionnaire was given to the participants who were then asked to log-in to the fictional website by writing down the password assigned to them. The objective password retention rate measure and the subjective satisfaction questionnaire indicated that all of the random generators produced passwords that were difficult for the users to remember. The DICEWARE group had the highest retention rate with two of seven participants recollecting their assigned password correctly. For both ALPHANUM and PRONOUNCE3 only one of six participants remembered their assigned passwords. The mean overall subjective satisfaction rating was 1.73 on a scale of 0-4, with 0 representing hate it and 4 love it. The subjective rating of the PRONOUNCE3 policy (mean = 1.83) was the highest followed by DICEWARE (mean = 1.71) and ALPHANUM (mean = 1.67).

Jeyaraman and Topkara (2005) developed a system that would generate a fictitious news headline as a mnemonic phrase to assist users in remembering their password. The system was tested with randomly generated lowercase passwords, for which it managed to create mnemonic headlines for 80.5% and 62.7% of six- and seven-character passwords respectively. The usability and user acceptance of the system was not evaluated.

These studies suggest that system-generated passwords are more secure but less usable than self-generated passwords. To address this issue, this study investigated the usability of a novel system-generated password with a mnemonic aid policy. While some researchers have used paper forms to represent computer systems for their usability studies, this study used a computer application to represent human interaction with computers more realistically. In addition, this study ensures a constant level of security or entropy among the four password generation policies investigated here. The entropy of the passwords generated by the four policies was 30±2 bits as recommended by NIST to attain its level 2 security standard. After the participants generate and log-in with their password, the NASA TLX measurement instrument was used to assess their cognitive and physical work load. Subjective satisfaction with each password generation policy was measured using a post-test System Usability Scale (SUS) questionnaire.

3. RESEARCH HYPOTHESES

The four password generation policies which were compared are

- **User-generated passwords with restrictions (User-generated password)**: In this policy the participants generated their own passwords following a set of instructions intended to prevent them from creating insecure passwords and ensuring minimum entropy of 30±2. The restrictions given to them in this case were that the password must be at least 8 characters long, contain at least one uppercase letter, one number and one special character. This password must also pass a dictionary check.
- **System-generated random passwords (System-generated password):** In this policy participants' were provided with a random 7 alphabetical character system-generated password having entropy of 30±2.
- **System-generated random password with mnemonic training (User-generated mnemonic):** In this policy users were provided with a system-generated password with 30±2 bit entropy just as in the previous condition. Participants were also provided with mnemonic aid generation training, and the mnemonic generated by them was collected.
- **System-generated random passwords with a system-generated mnemonic aid (System-generated mnemonic):** In this policy the participants were provided with a random password as in the previous two

conditions and with a system-generated mnemonic aid. For example, if the system generated password was **vpgbeii**, **Victor's pet goat briefly examined individual insects**, was provided as a system-generated mnemonic aid.

To compare the usability of these policies, the following research hypotheses were investigated.

Hypothesis 1:

It is hypothesized that in terms of user satisfaction

the system-generated password and mnemonic aid will be at least as satisfactory as the user-generated password with restrictions and the system-generated password with mnemonic generation training but more satisfactory than the system-generated random password.

It is expected that the system-generated password linked with a system-generated mnemonic will be easier for the users to remember than the system-generated random password alone. Thus, system-generated passwords with a system-generated mnemonic users are expected be more satisfied than system-generated password users.

Hypothesis 2:

It is hypothesized that in terms of password retention accuracy the password retention accuracy of the system-generated password linked to a mnemonic aid will be at least equal to that of the user-generated password with restrictions and the system-generated password with mnemonic generation training and higher than the system-generated password.

It is expected that the system-generated password linked with a system-generated mnemonic will help the users to more accurately recollect their passwords than the system-generated passwords.

Hypothesis 3:

It is hypothesized that in terms of workload:

the system-generated password linked with a system-generated mnemonic will result in less workload than the user-generated password with restrictions, the system-generated password with mnemonic generation training and the computer-generated password policies.

It is expected that the system-generated password linked with a system-generated mnemonic will help the users to generate their passwords as well as to memorize them with less effort than the system-generated password, the system-generated password with mnemonic generation training and the user-generated password with restrictions.

Hypothesis 4:

It is hypothesized that in terms of the time required to create and memorize the passwords

the time taken by the participants to successfully enter the system-generated password linked to a system-generated mnemonic will be less than the system-generated password with mnemonic generation training, and approximately equal to the system-generated password and the user-generated password with restrictions.

It is expected that the system-generated password linked with a system-generated mnemonic will help the users to quickly create and remember their password and will also enable them to complete their password creation and log-in tasks faster than the system-generated password with mnemonic generation training.

Hypothesis 5:

It is hypothesized that in terms of the number of errors made by the participants while creating/recalling the passwords

the total number of errors made by the participants while creating/recalling the system-generated password linked to a system-generated mnemonic will be less than the system-generated password, the system-generated password with mnemonic generation training and the user-generated password with restrictions.

It is expected that the system-generated password linked with a system-generated mnemonic will help the users to create and remember their passwords correctly with fewer errors.

4. EXPERIMENTAL METHOD

Participants

Sixty-four students from Clemson University were recruited through an email and/or verbal invitation describing this study. Students expressing an interest in participating were pre-screened via questionnaire to determine their eligibility: participants were required to have prior experience using the Internet for a minimum of one year. In addition, they were required to have experience in constructing and maintaining passwords for user accounts on the Web. This pool of 64 participants was randomly divided into four groups: 16 in Group 1 representing user-generated password with restrictions, 16 in Group 2 representing system-generated passwords, 16 in Group 3 representing system-generated passwords with mnemonics creation training for the users, and 16 in Group 4 representing system-generated passwords and mnemonics.

Experimental Design

This experiment is considered to be both a one-factor design with four levels and a two-factor design with four levels of the first factors two levels and two levels of the second factor. The independent variable of the former investigates the password composition scheme at the four levels defined in Table 4.1. Each of the four conditions, or levels, of the independent variable, password construction policy, used the same minimum password guessing entropy of 30±2 bits. The assignment of participants to these conditions was random, subject to the constraint that an equal number of participants were assigned to each. The data was collected from each participant over two sessions and subsequently statistically analyzed.

Table 4.1: One Factor design with four levels

Level 1 **Scheme**	**Level 2** **Scheme**	**Level 3** **Scheme**	**Level 4** **Scheme**
User-generated password with restrictions.	System-generated random password	System-generated random password with mnemonic generation training for the participants	System-generated random password and mnemonic
Minimum of 8 characters.	7 characters	7 characters	7 characters
At least one lower and one upper case letter, one number and one special character.	Random characters selected from any of the 26 lower case letters available on the standard QWERTY keyboard	Random characters selected from any of the 26 lower case letters available on the standard QWERTY keyboard	Random characters selected from any of the 26 lower case letters available on the standard QWERTY keyboard
No common words or character sequences or permutations of usernames.	No common words or character sequences or permutations of usernames.	No common words or character sequences or permutations of usernames.	No common words or character sequences or permutations of usernames.

For the dependent variables recorded in both of the recall task sessions, the experiment was a two-factor design with four levels of password composition scheme and two levels of recall task session. The second independent variable of the study was the recall task sessions defined in Table 4.2.

Table 4.2: 4x2 factorial design

IVs	**Session 1: Recall**	**Session 2: Recall**
Password composition scheme: Condition 1		
Password composition scheme: Condition 2		
Password composition scheme: Condition 3		
Password composition scheme: Condition 4		

The dependent variables in this experiment include objective and subjective measures of performance. The experimental study was conducted in two sessions, the first one in which the participants created and/or memorized their password, recalled their password after a five minute distraction task and the second in which they recalled them after a week's time. The objective measures for the first session are the number of password creation/recall errors made by the participants and the total time taken to create and memorize their passwords. The objective measures for the second session are the number of password recall errors and the total time taken by the participants to recall and enter their passwords after a 5 to 7 day interval. Password retention accuracy was also measured for the recall task in both sessions, using the Damerau–Levenshtein edit distance (Damerau, 1964) and Jaro-Winkler proximity edit distance (Winkler, 1990). The Damerau–Levenshtein edit distance between two strings is defined as the minimum number of edits, i.e. total sum of single character insertions, deletions, substitutions, and adjacent transpositions needed to transform a recalled password into the actual one. The Jaro-Winkler proximity edit distance between two strings is the similarity or correlation between the recalled password and the actual password stored in the first session, normalized such that 0 indicates no similarity and 1 indicates equality.

Subjective data were obtained using the System Usability Scale (SUS) questionnaire (See Appendix D) administered to the participants at the end of each task in both sessions of the experimental study. The questionnaire at the end of the first session creation and/or memorization task addressed the ease of creating/memorizing the password and the questionnaire at the end of first session recall task addressed the ease of recalling the passwords for this session. The questionnaire administered to the participants at the end of the second session addressed the long term memorability of the passwords created/memorized. In addition, at the end of each task, the NASA TLX workload questionnaire (See Appendix E) was administered to the participants to measure perceived workload.

Testing Environment

The study was conducted in the Human Computer Systems Laboratory at Clemson University. The experimental set-up consisted of a desktop computer, table, chair, paper and pencil. The computer screen displayed a password log-in application for which participants either created a password or were assigned a system-generated one. This application provided immediate feedback on whether the password created conformed to the stipulated password policies/guidelines before accepting it.

Tasks

The experimental study was conducted over two sessions, the first lasting approximately 15 minutes and the second lasting approximately 5 minutes with a 5 to 7 day interval between them. In the first session, the participants created and/or memorized their passwords as explained below:

1. All participants: were assigned a log-in user name.
2. Group 1 Participants: Created a password following the instructions provided as shown in Figure 4.1. Used this password to log in to the application.

   Groups 2 Participants: Memorized a system-generated password. See Figure 4.2. Used this password to log in to the application.

   Group 3 Participants: Created a mnemonic aid for the system-generated password assigned to them based on the training provided and memorized the password. See Figure 4.3. Used this password to log in to the application.

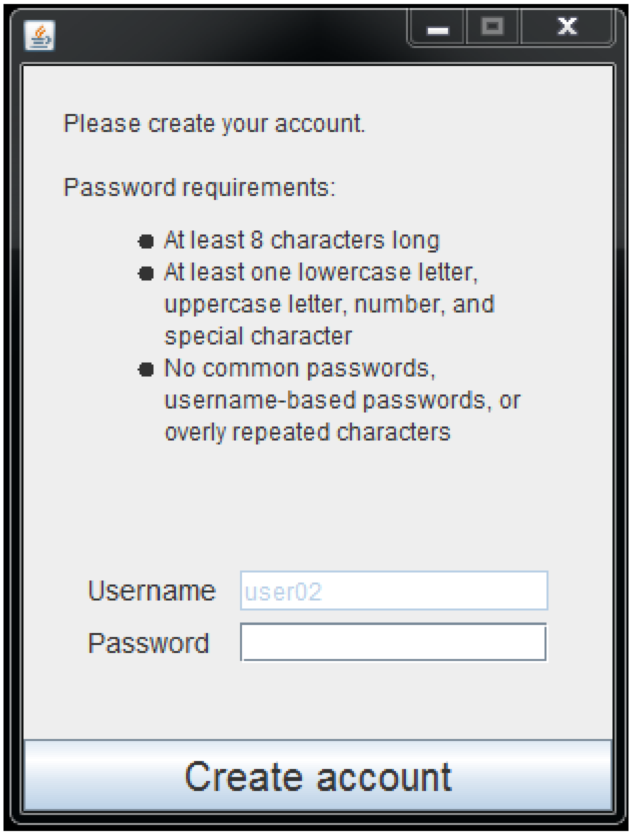
   Group 4 Participants: Used the system-generated mnemonic aid to memorize the system-generated password assigned to them. See Figure 4.4. Used this password to log in to the application.
3. All participants checked the feedback provided by the password log-in application.

   If the feedback indicated that the password did not conform to its requirements, Group 1 participants were again asked to create a new password conforming to the instructions provided to them. All the other group participants were shown the system-generated password assigned to them in Step 2. See Figures 4.5, 4.6, 4.7, 4.8.

4. If the password entered was correct, the participants were asked to complete the NASA TLX work load assessment and the System Usability Scale (SUS) questionnaire. Then they were asked to perform a distraction task of playing the Angry Birds© game (Lehtinen, 2009) for 5 minutes.
5. After completing the distraction task, the participants were asked to log in using their assigned or created passwords. A total of five attempts were permitted to enter the password correctly for the first time. See Figures 4.9 and 4.10.
6. All participants completed the NASA TLX work load assessment.
7. All participants completed the System Usability Scale (SUS).

The participants were then asked to return 5 to 7 days later, depending on their availability, to perform the following tasks:

1. All participants entered their previously assigned or created password into the login application with a total of five attempts permitted to enter the password correctly for the first time.
2. All participants completed the NASA TLX.
3. All participants completed the System Usability Scale (SUS).



Please create your account.

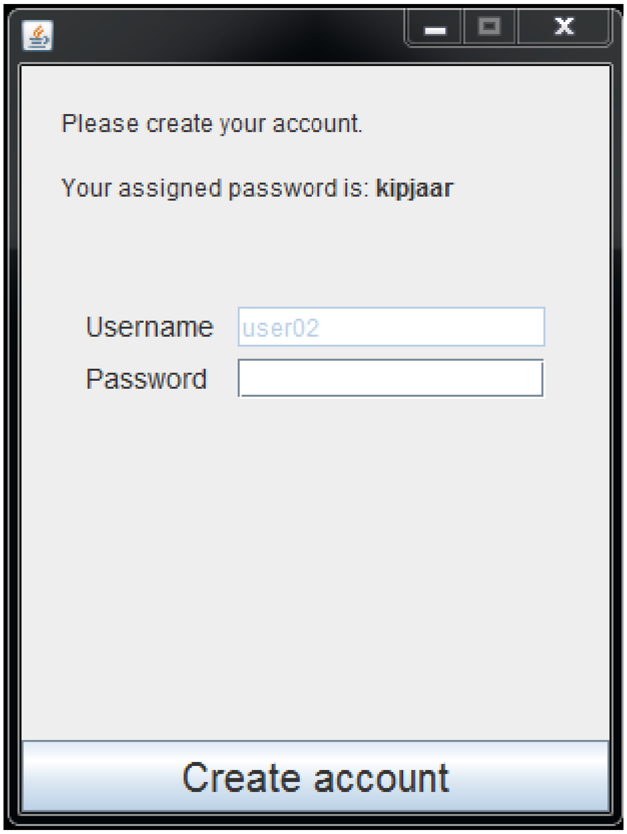
Password requirements:

- At least 8 characters long
- At least one lowercase letter, uppercase letter, number, and special character
- No common passwords, username-based passwords, or overly repeated characters

Username user02

Password

Figure 4.1: 8-character user-generated password creation



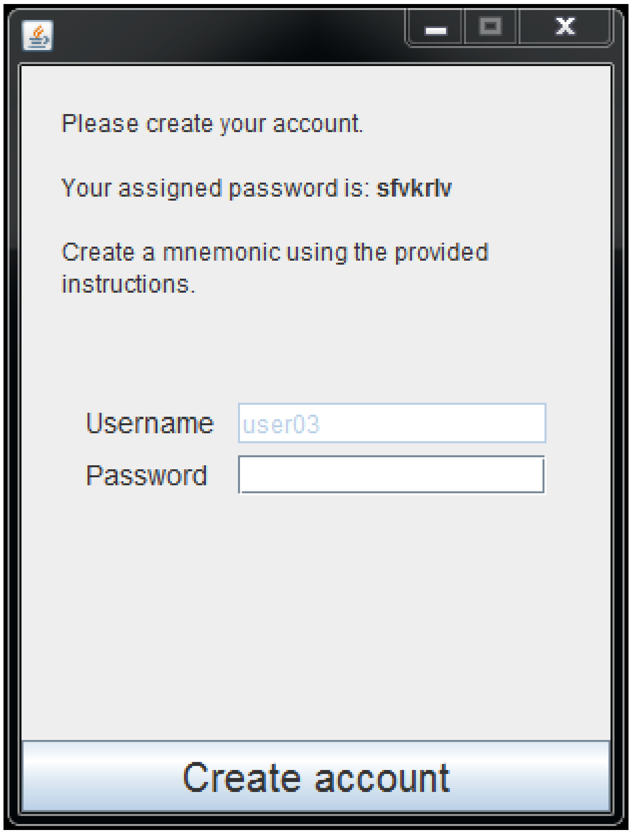
Please create your account.

Your assigned password is: **kipjaar**

Username: user02

Password:

Figure 4.2: 7-character system-generated password creation



Please create your account.

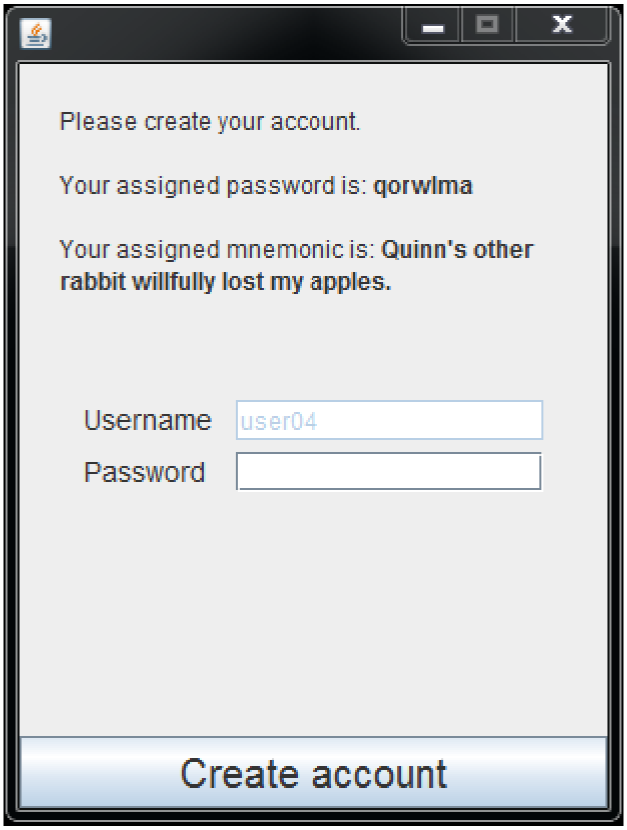
Your assigned password is: **sfvkrlv**

Create a mnemonic using the provided instructions.

Username: user03

Password:

Figure 4.3: 7-character system generated password and user-generated mnemonic creation



Please create your account.

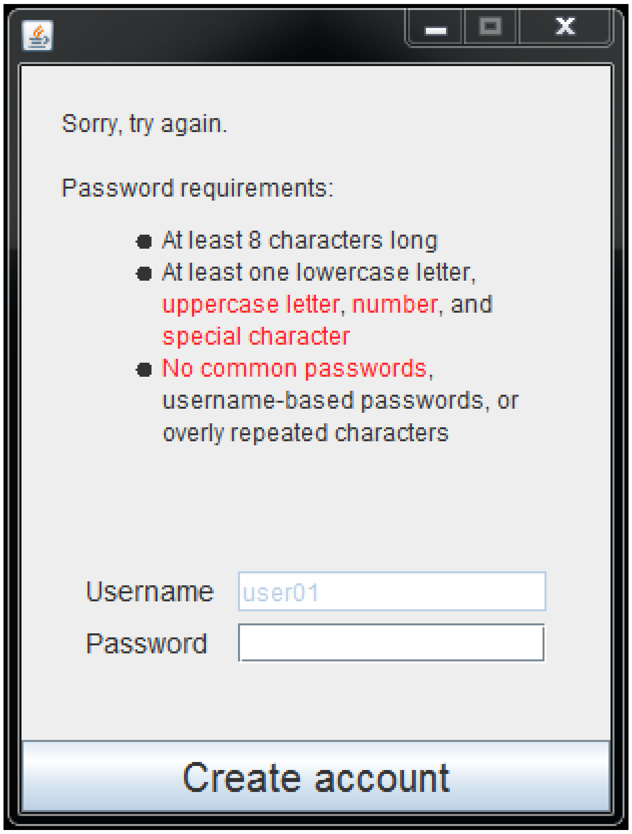
Your assigned password is: **qorwlma**

Your assigned mnemonic is: **Quinn's other rabbit willfully lost my apples.**

Username: user04

Password:

Figure 4.4: 7-character system-generated password and mnemonic creation



Sorry, try again.

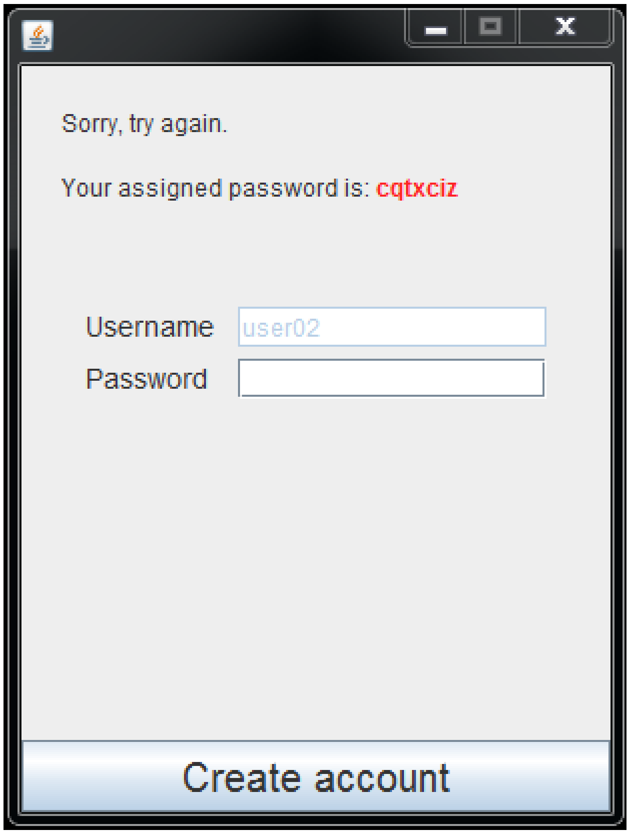
Password requirements:

- At least 8 characters long
- At least one lowercase letter, uppercase letter, number, and special character
- No common passwords, username-based passwords, or overly repeated characters

Username: user01

Password:

Figure 4.5: Response popup window to a failed 8-character user-generated password creation



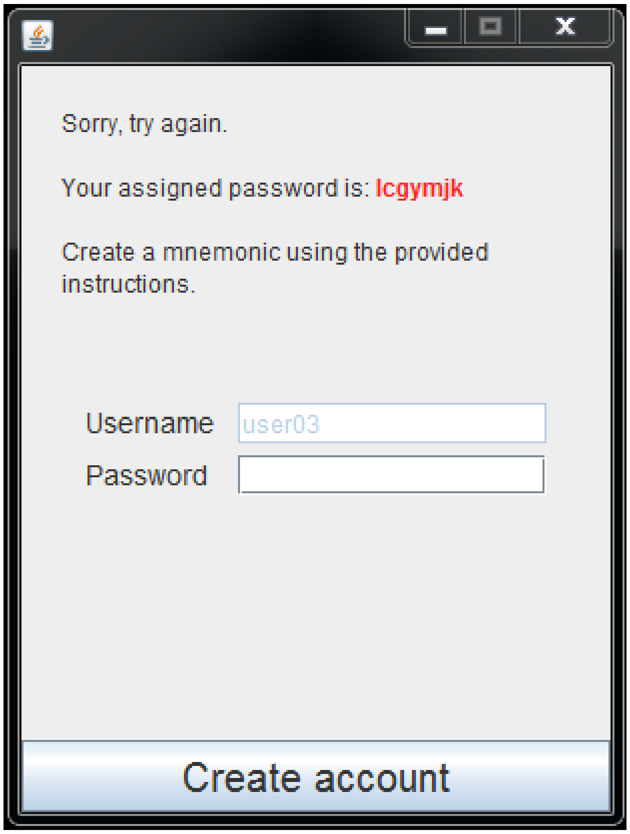
Sorry, try again.

Your assigned password is: cqtxciz

Username: user02

Password:

Figure 4.6: Response popup window to a failed 7-character system-generated password creation



Sorry, try again.

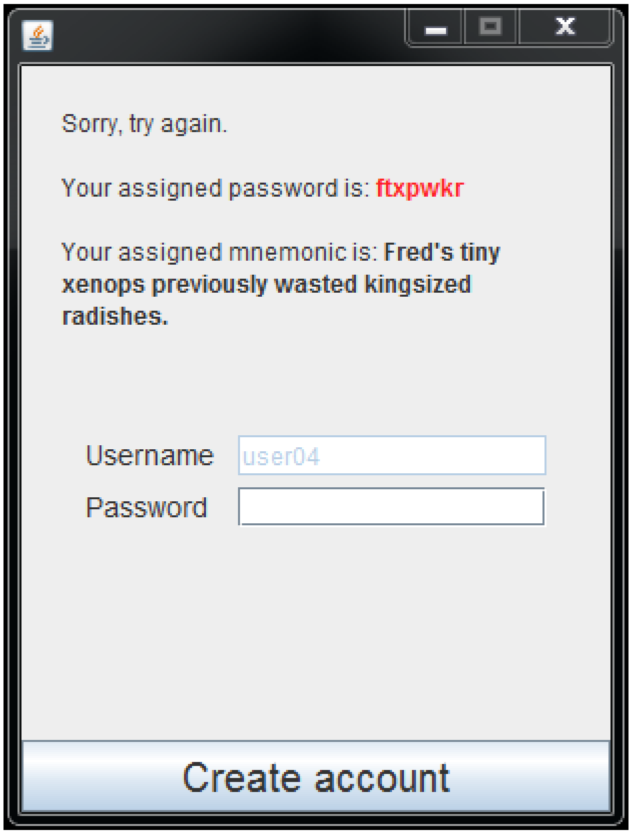
Your assigned password is: lcgymjk

Create a mnemonic using the provided instructions.

Username user03

Password

Figure 4.7: Response popup window to a failed 7-character system generated password and user-generated mnemonic creation



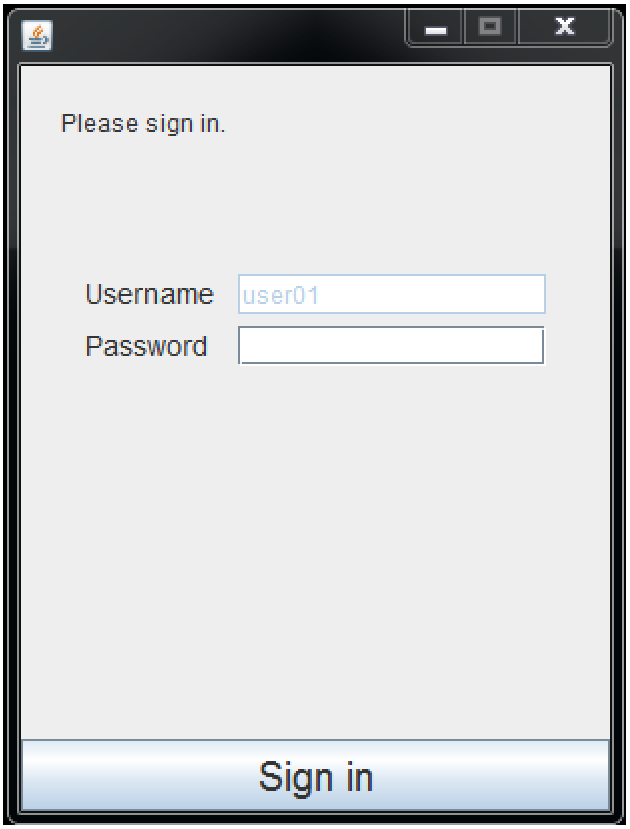
Sorry, try again.

Your assigned password is: **ftxpwkr**

Your assigned mnemonic is: **Fred's tiny xenops previously wasted kingsized radishes.**

Username: user04

Figure 4.8: Response popup window to a failed 7-character system-generated password and mnemonic creation

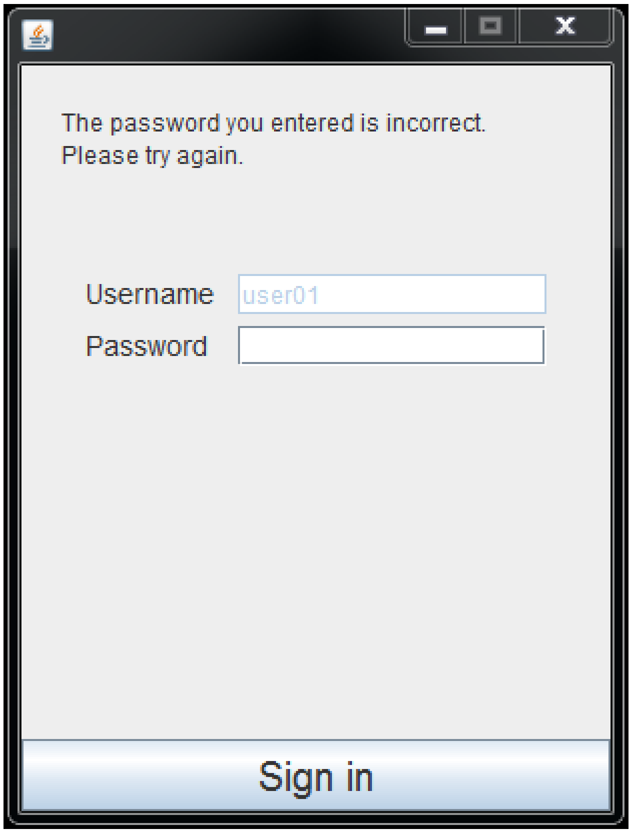


Please sign in.

Username user01

Password

Figure 4.9: Password recall pop-up window



The password you entered is incorrect.
Please try again.

Username user01

Password

Figure 4.10: Failed password recall attempt

Procedure

At the beginning of the first session, the researcher greeted the participant, who was then seated in front of a desktop computer on a table in the Human Computer Systems Laboratory. The researcher provided a brief overview of the experiment to the participant. After the participant read and signed the informed consent form (See Appendix A), they completed a pre-study questionnaire (See Appendix B) asking for demographics, information on their Internet experience and their previous experience in creating user accounts on the Internet. After completion of the pre-study questionnaire, the researcher provided training on the types of passwords that were not accepted by a dictionary check for Condition 1 participants and memory tools such as mnemonics for Condition 3 participants (See Appendix C). The duration of this training was approximately 5 minutes.

After the completion of training, the participant either created their password or memorized their assigned password conforming to the password guidelines provided and subsequently entered the password into the password log-in application on the desktop computer. The application provided immediate feedback regarding the acceptability of the password. For the user-generated password condition, the application provided feedback on the conformation of the password created to the required guidelines, failing which the participant was asked to create a new password. The time taken and the number of errors committed during the entry of passwords in the first session were recorded. After a five-minute distraction task of playing the Angry Birds$^{©}$ game (Lehtinen, 2009), the participant again entered the password created or assigned into the application, with five attempts being allowed to make a correct entry. The time taken to enter the correct password and the log-in error rate were recorded.

On completion of each of the above creation and/or memorization task and the password recollection task after a 5 minute distraction, the participant was asked to complete the NASA Task Load Index questionnaire (See Appendix D) to assess the perceived workload experienced during those tasks. Then, the participant was administered the System Usability Scale (SUS) questionnaire (See Appendix E). These questions used a 5-point Likert scale, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). At the end of the session, the researcher asked the participant to schedule a date and time for the second session of the experimental study. The participant was also asked to try to remember the password they had created or been assigned as well as to avoid externalizing it. The duration of the first phase of the study was approximately 15 minutes.

At the beginning of the second session, the researcher briefed the participant on the task to be conducted. The researcher asked the participant to recall their password from the first session and to enter it into the password login application on the desktop computer. The time taken to make the first successful login was recorded. A maximum of five attempts was given to the participant to recall his or her password correctly; if the participant failed to be able to do so, the password was specified as unrecoverable. The smallest Damerau–Levenshtein edit distance number and the greatest Jaro-Winkler proximity edit distance number obtained in the five unsuccessful attempts was recorded.

The participant was asked to complete the NASA Task Load Index questionnaire to assess the perceived workload experienced during the login task (See Appendix D). The researcher then administered the System Usability Scale questionnaire (See Appendix E) to the participant. The duration of the second phase was approximately 5 minutes. See the procedure flow for the 1st and 2nd sessions in Figure 4.11.

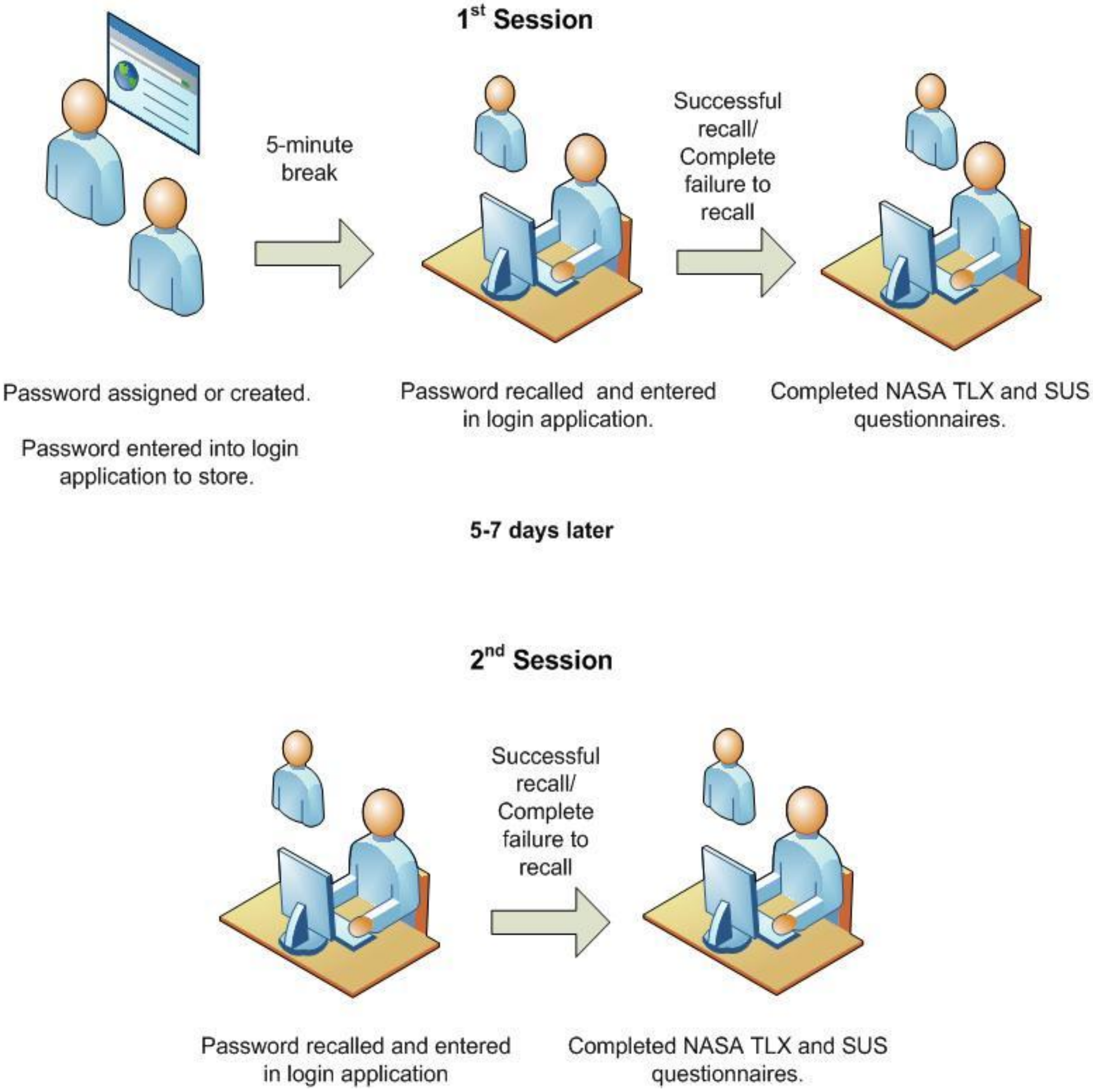

1st Session
5-minute break
Successful recall/ Complete failure to recall
Password assigned or created.
Password entered into login application to store.
Password recalled and entered in login application.
Completed NASA TLX and SUS questionnaires.
5-7 days later
2nd Session
Successful recall/ Complete failure to recall
Password recalled and entered in login application
Completed NASA TLX and SUS questionnaires.

Figure 4.11: Procedure flow for 1st and 2nd session

5. RESULTS

The first and second sessions of the experiment were completed by 64 out of 73 participants. Nine participants failed the recall task in the first session. Their data were not included in the complete statistical analysis. The reasons for their failure were analyzed separately through quantitative and qualitative data collected from them. The statistical analysis software SPSS 19 was used for data analysis. The data collected across task sessions from all the participants were checked for normality. These results showed that the dependent measures in the first session, password creation and/or memorization time and error rate, were non-normal. In both sessions, password recall times, recall error rates, and the edit distances (Damerau-Levenshtein and Jaro-Winkler) were non-normal, exhibiting high skewness values. The data from these dependent measures were transformed using the reciprocal function to normalize them. Even after this transformation, the recall error rate and edit distance data across sessions were not normally distributed. As a result, these measures were analyzed using non-parametric tests.

The dependent measures of recall time, recall NASA TLX work load measure and recall System usability scale (SUS) were measured twice over the task sessions with the same participants, after which they were analyzed for significance using repeated measures two-way ANOVA with a 95% confidence interval. The dependent measures of password creation/memorization time, password creation error rate, and creation/memorization NASA TLX work load were only measured once and were, thus, analyzed for significance across conditions using one-way ANOVA with a 95% confidence interval. Then the locus of the significance, if any, was determined using the Least Significant Difference (LSD) post-hoc test.

Objective measures

The objective measures recorded in the first session involved the creation/memorization task and the recall of the password after a 5 minute interval. These measures for the first task consisted of the creation/memorization time and creation error rate, the recall time, recall error rate, and the recall edit distance for the recall task after 5 minutes. In the second session in which the participants recalled their passwords after a week, the objective measures recorded were recall time, recall error rate and recall edit distance.

Creation/Memorization Time

The creation/memorization time which was used to determine the creation efficiency of password policies, includes the time taken by the participants to create and/or memorize a password based on the assigned policy and to type it into the system and successfully create an account. The data collected from the 64 participants were statistically analyzed, the results indicating they were not normally distributed. As a result, the data were transformed into their inverse to normalize them and then analyzed again. The descriptive statistics of this inversed data are provided in Table 5.1, the numbers in the parentheses representing the actual mean creation time measured in seconds.

Table 5.1: Descriptive statistics for the time taken to create password accounts

	N	Inverse Mean(Actual Mean time in Seconds)	Inverse data Std. Deviation	Inverse data Std. Error
User-Generated	16	.0222 (61.97)	.01162	.00291
System-Generated	16	.0255(54.43)	.01593	.00398
User-Generated Mnemonic	16	.0075(207.13)	.00423	.00106
System-Generated Mnemonic	16	.0207(70.52)	.01510	.00377
Total	64	.0190(98.51)	.01409	.00176

The analysis conducted using one-way ANOVA found a significant difference (F (3,63)=6.289, p = 0.001) across the conditions as shown in Table 5.2.

Table 5.2: One-way ANOVA data for time taken to create password accounts

Creation/Memorization Time	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.003	3	.001	6.289	.001
Within Groups	.010	60	.000		
Total	.013	63			

The subsequent LSD post-hoc test indicated the password created using the user-generated mnemonic policy (Condition 3) took significantly more time to memorize than the other three conditions. These conditions did not exhibit a significantly different creation/memorization time among them. Figure 5.1 and Figure 5.2 present the mean actual creation/memorization time and the transposed creation/memorization time (1-

actual time^{-1}) across, the conditions, respectively. In order to maintain the nature (slope) of the graph we transpose the inversed data by subtracting it from one.

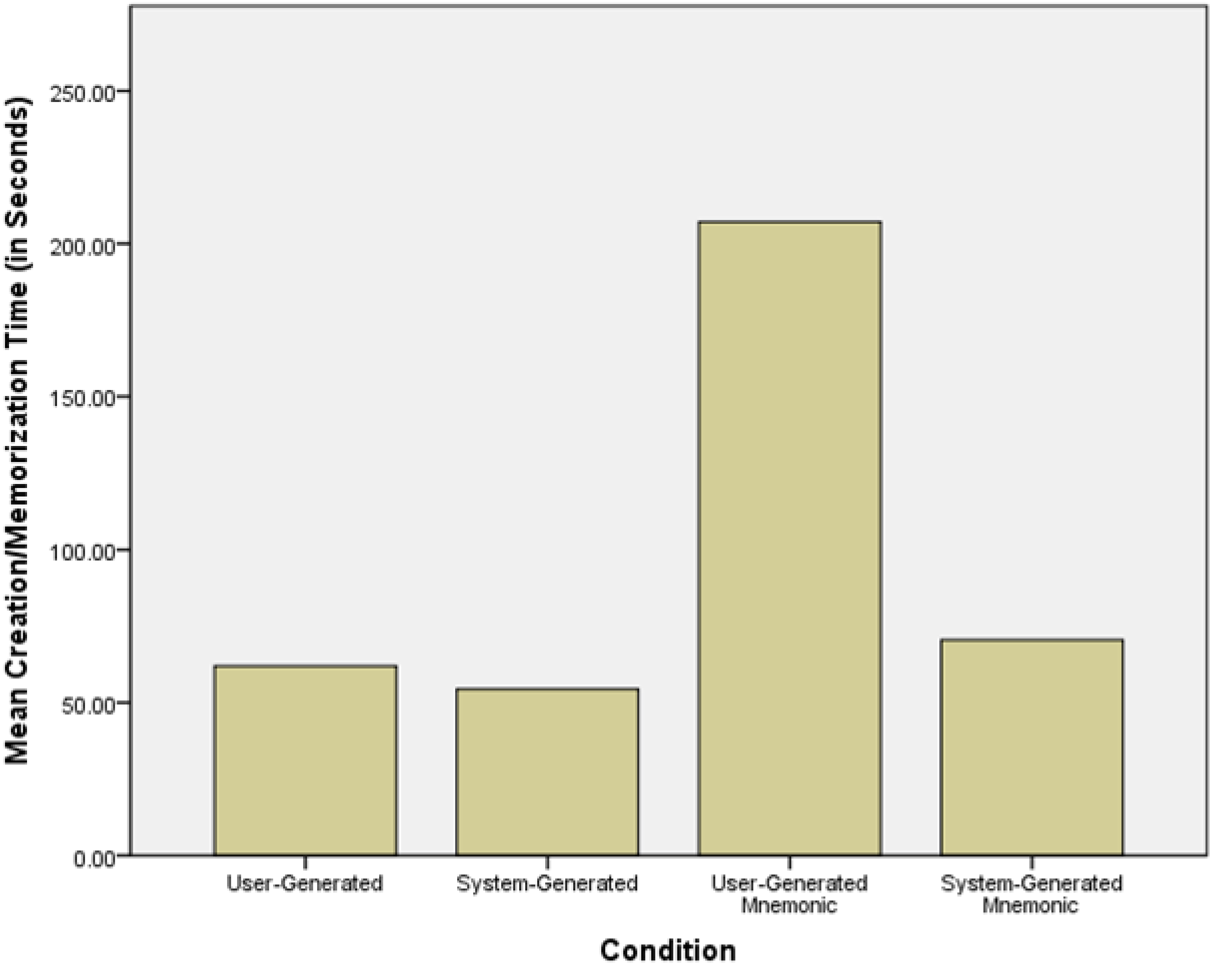


Figure 5.1: Mean time taken (in Seconds) to create/memorize the password

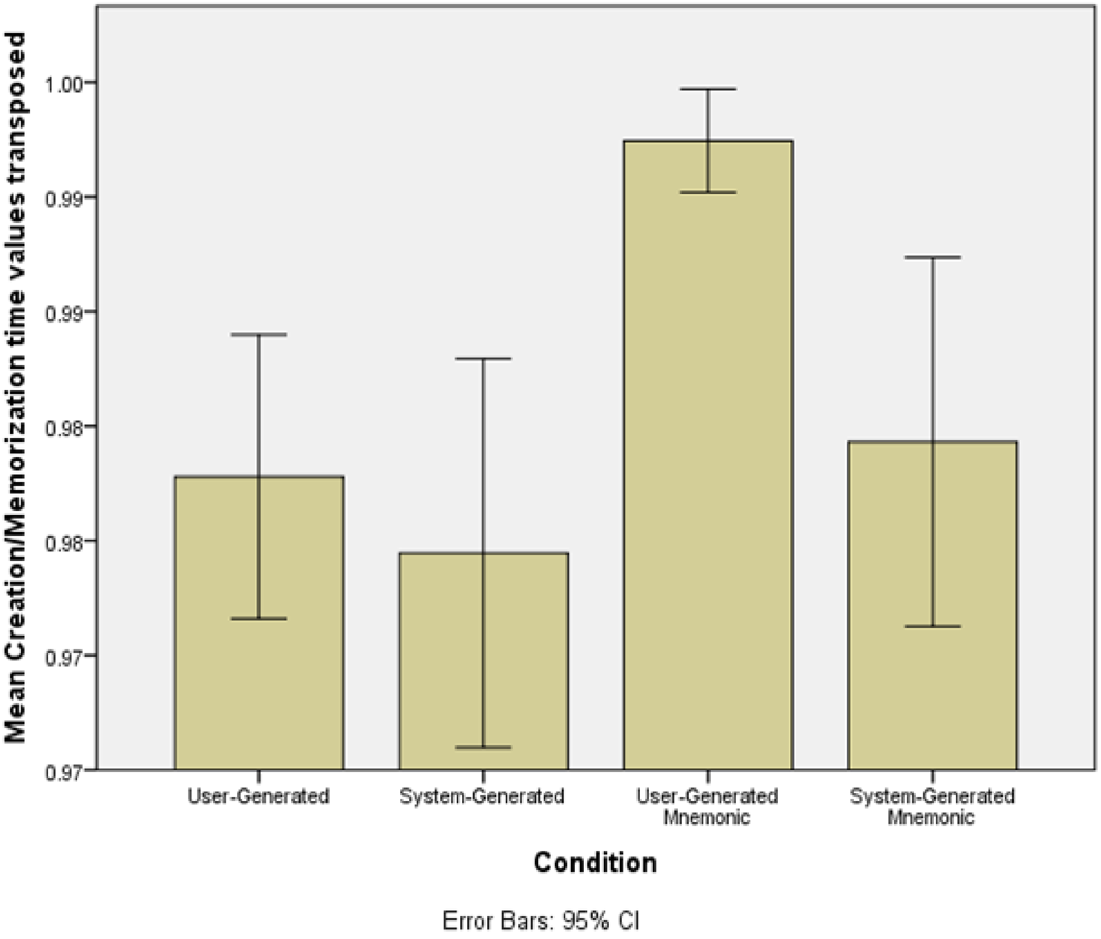


Figure 5.2: Mean transposed time (1-actual time^{-1}) taken to create a password account

Creation Error Rate

The creation error rate was used to determine the creation effectiveness of the password policies. This metric was measured by dividing the number of errors by the total number of attempts taken to create the password account. Error rate was used instead of error count because it would be a more holistic measure and easy to compare among groups. The data analysis showed that this dependent variable was not normal even after the data were subjected to inverse transformation; the descriptive statistics for this metric are provided in Table 5.3. The Kruskal-Wallis non-parametric test was used for further analysis, the results of which are shown in Table 5.4. As this table indicates, there is no significant difference among various password policies (H(3)=3.709,p=0.295).

Table 5.3: Descriptive statistics for error rates during password account creation

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	.0625	.17078	.04270
System-Generated	16	.1042	.22675	.05669
User-Generated Mnemonic	16	.0313	.12500	.03125
System-Generated Mnemonic	16	.0000	.00000	.00000
Total	64	.0495	.15625	.01953

Table 5.4: Kruskal-Wallis test on the password account creation error rate

	Condition	N	Mean Rank
Creation/Memorization Error Rate	User-Generated	16	33.44
	System-Generated	16	35.59
	User-Generated Mnemonic	16	31.47
	System-Generated Mnemonic	16	29.50
	Total	64	

Test Statistics[a,b]

	Creation/Memorization Error Rate
Chi-Square	3.709
kdf	3
Asymp. Sig.	.295

a. Kruskal-Wallis Test

b. Grouping Variable: Condition

Recall Time

The system recorded the time taken by the participants to recall their passwords and enter them after a five-minute distraction task in the first session and after a week in the second session. Since the data collected were not normal for both the sessions, the log transformation was applied to normalize them. The descriptive statistics for this metric for both sessions are provided in Tables 5.5 and 5.6, the mean, standard deviation and error indicating the logarithmic values and the numbers in the parentheses representing the actual mean recall time in seconds.

Table 5.5: Descriptive statistics for recall times for first session

Password Policies	N	Log transformation of the Mean time (Actual Mean time in Seconds)	Logarithmic data Std. Deviation	Logarithmic data Std. Error
User-Generated	16	2.36(11.28)	.34727	.08682
System-Generated	16	2.28(10.73)	.40875	.10219
User-Generated Mnemonic	16	2.57(15.33)	.55026	.13756
System-Generated Mnemonic	16	2.31(14.45)	.82888	.20722
Total	64	2.38(12.95)	.56276	.07035

Table 5.6: Descriptive statistics for recall times for second session

Password Policies	N	Log transformation of the Mean time(Actual Mean time in Seconds)	Logarithmic data Std. Deviation	Logarithmic data Std. Error
User-Generated	16	2.45(17.84)	.72251	.18063
System-Generated	16	3.02(50.65)	1.37483	.34371
User-Generated Mnemonic	16	3.28(54.44)	1.11950	.27987
System-Generated Mnemonic	16	2.91(52.88)	1.54181	.38545
Total	64	2.91(43.95)	1.23678	.15460

One-way ANOVA was used to analyze the data for significant differences among password policies, the results indicating no significant difference in recall time among the password policies in either session (Session 1; F(3,63)=0.824, p=0.486, Session 2; F(3,63)=1.264, p=0.295) as shown in Table 5.7 and 5.8.

Table 5.7: One-way ANOVA data for time taken to recall password in 1st session

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.790	3	.263	0.824	0.486
Within Groups	19.163	60	.319		
Total	19.952	63			

Table 5.8: One-way ANOVA data for time taken to recall password in 2nd session

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5.726	3	1.909	1.264	0.295
Within Groups	90.639	60	1.511		
Total	96.366	63			

A two-way mixed ANOVA was also conducted to test the main and interaction effects of the password policy conditions and the task sessions on the time taken to recall the passwords. The result indicated that the main effect was significant for task session, F (1, 60) =4.369, *p*=0.041 but not significant for password creation condition F(3,60)=2.134, *p*=0.105. The two-way ANOVA data for the transposed value of the recall times are provided in Table 5.9. Subsequent post-hoc analysis of the task session main effect revealed that the time taken to recall a password was less for the first session than for the second (*p*=0.041). The interaction effect of password policy conditions and task sessions on the time taken to recall passwords was not significant, F (3, 60) =0.742, p=0.531. The interaction effects of the inversed recall time, transposed recall time and the actual recall time are plotted in Figure 5.3(a), 5.3(b) and Figure 5.4, respectively.

Table 5.9: Two-way ANOVA data for recall times

Recall Times	SS	df	Mean Squares	F	Sig.
Task Sessions	1.722	1	1.722	15.922	0.000
Conditions	0.039	3	.013	2.134	0.105
Task Sessions x Conditions	0.408	3	0.136	1.258	0.297
Error (Within-subjects)	0.100	60	.002		
Error (Between-subjects)	0.362	3	.006		

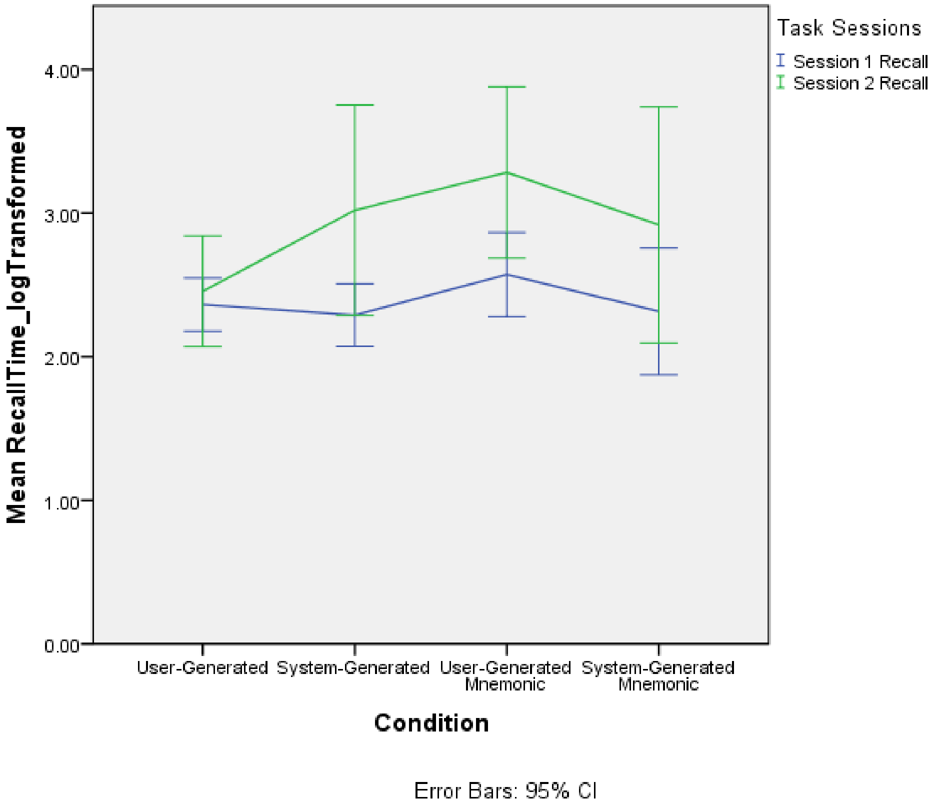


Figure 5.3: Interaction effect plots of the log transformation of time taken to recall password

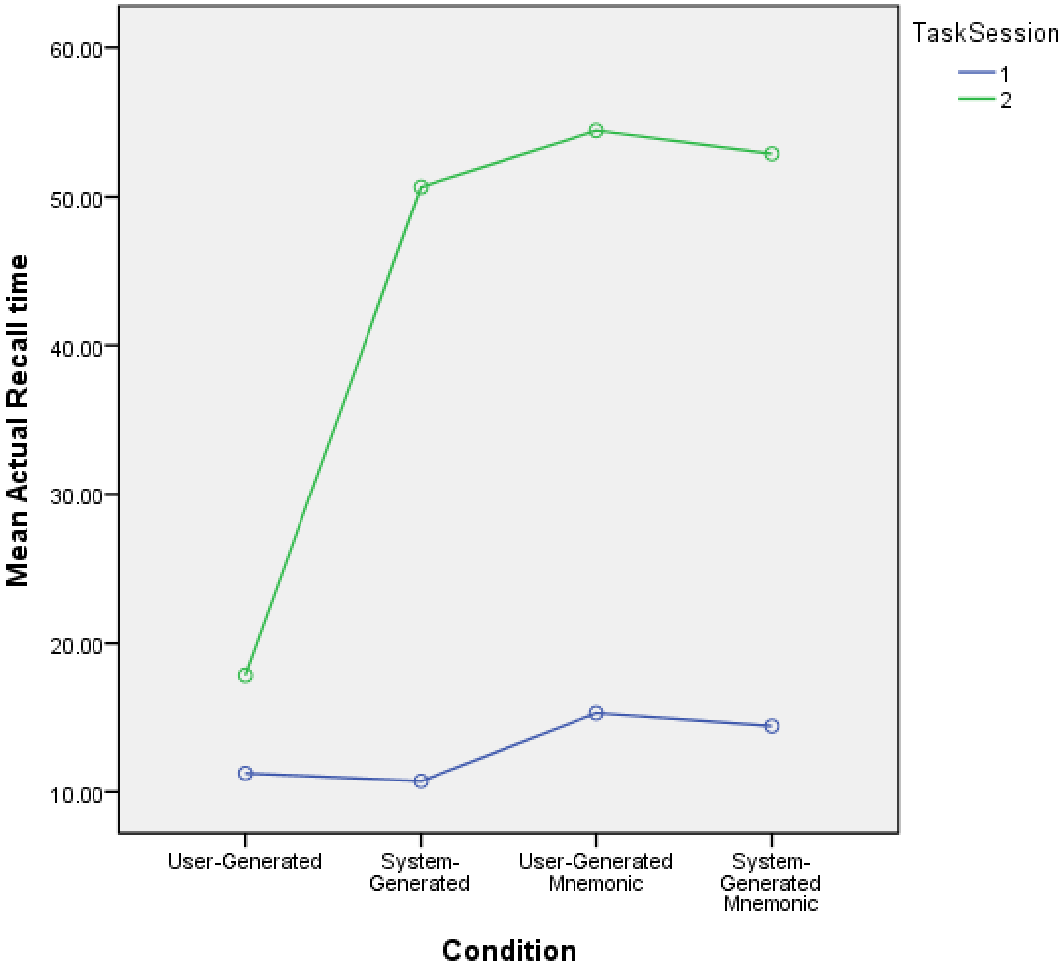


Figure 5.4: Interaction effect plots of the actual values of time taken (in Seconds) to recall password

Recall Error Rate

The recall error rate measures the ratio between the total number of failed attempts to enter the correct password and the total attempts taken to enter the correct password. This measure helps to determine how effectively people remembered and recollected their password in both the sessions. The data analysis from both showed that this dependent variable was not normal even after the data were subjected to inverse transformation. The descriptive statistics for this metric are provided in Tables 5.10 and 5.11. Further statistical analysis was conducted using the Kruskal-Wallis non-parametric test, the results for each session being shown in Tables 5.12 and 5.13. These tables show that there are no significant differences among the password policies for either session (Session 1, H (3) =1.350, p =0 .717; Session 2, H (3) =1.306, p=0.728).

Table 5.10: Descriptive statistics for error rates during 1st session password recall

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	.0625	.17078	.04270
System-Generated	16	.0313	.12500	.03125
User-Generated Mnemonic	16	.0625	.17078	.04270
System-Generated Mnemonic	16	.1146	.24894	.06223
Total	64	.0677	.18241	.02280

Table 5.11: Descriptive statistics for error rates during 2nd session password recall

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	.1563	.30104	.07526
System-Generated	16	.3125	.47871	.11968
User-Generated Mnemonic	16	.2969	.42050	.10513
System-Generated Mnemonic	16	.3438	.47324	.11831
Total	64	.2773	.42050	.05256

Table 5.12: Kruskal-Wallis test on the password account 1st session recall error rate

	Condition	N	Mean Rank
Recall Error Rate	User-Generated	16	32.38
	System-Generated	16	30.44
	User-Generated Mnemonic	16	32.38
	System-Generated Mnemonic	16	34.81
	Total	64	

Test Statistics[a,b]

	Recall Error Rate
Chi-Square	1.350
df	3
Asymp. Sig.	.717

a. Kruskal-Wallis Test

b. Grouping Variable: Condition

Table 5.13: Kruskal-Wallis test on the password account 2nd session recall error rate

	Condition	N	Mean Rank
Recall Error Rate	User-Generated	16	28.81
	System-Generated	16	33.09
	User-Generated Mnemonic	16	33.47
	System-Generated Mnemonic	16	34.63
	Total	64	

Test Statistics[a,b]

	Recall Error Rate
Chi-Square	1.306
df	3
Asymp. Sig.	.728

a. Kruskal-Wallis Test

b. Grouping Variable: Condition

A Friedman non-parametric test was also conducted to examine the main effect of task session on the error rate in recalling the passwords, the results indicating that the main effect was significant, χ^2 (1) =9.846, *p*=0.002. The Friedman's test results for the error rates are provided in Table 5.14. The error rate for recalling a password was lower for the first session than for the second (*p*=0.001).

Table 5.14: Friedman test on the password account recall error rate

	Mean Rank
Session 1 Recall Error Rate	1.38
Session 2 Recall Error Rate	1.63

Test Statistics[a]

N	64
Chi-Square	9.846
df	1
Asymp. Sig.	.002

a. Friedman Test

Unrecoverable passwords

One user-generated, three system-generated, one user-generated mnemonic and four system-generated mnemonic condition participants failed to recall their passwords in the first session as shown in Figure 5.5. In the second session one user-generated, five system-generated, three user-generated mnemonic and five system-generated mnemonic condition participants failed to recall their passwords as shown in Figure 5.6.

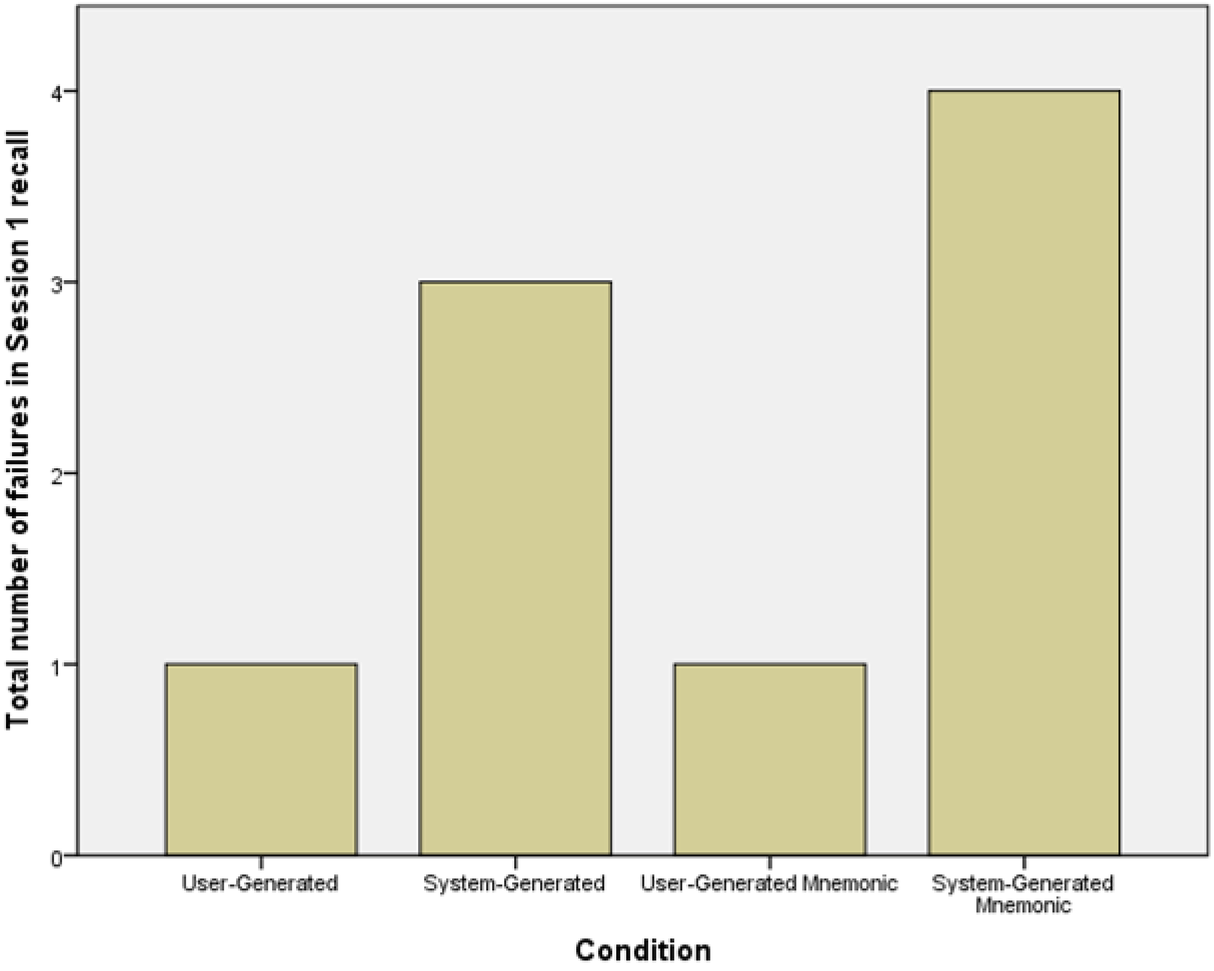


Figure 5.5: Distribution of the participants failing in session 1

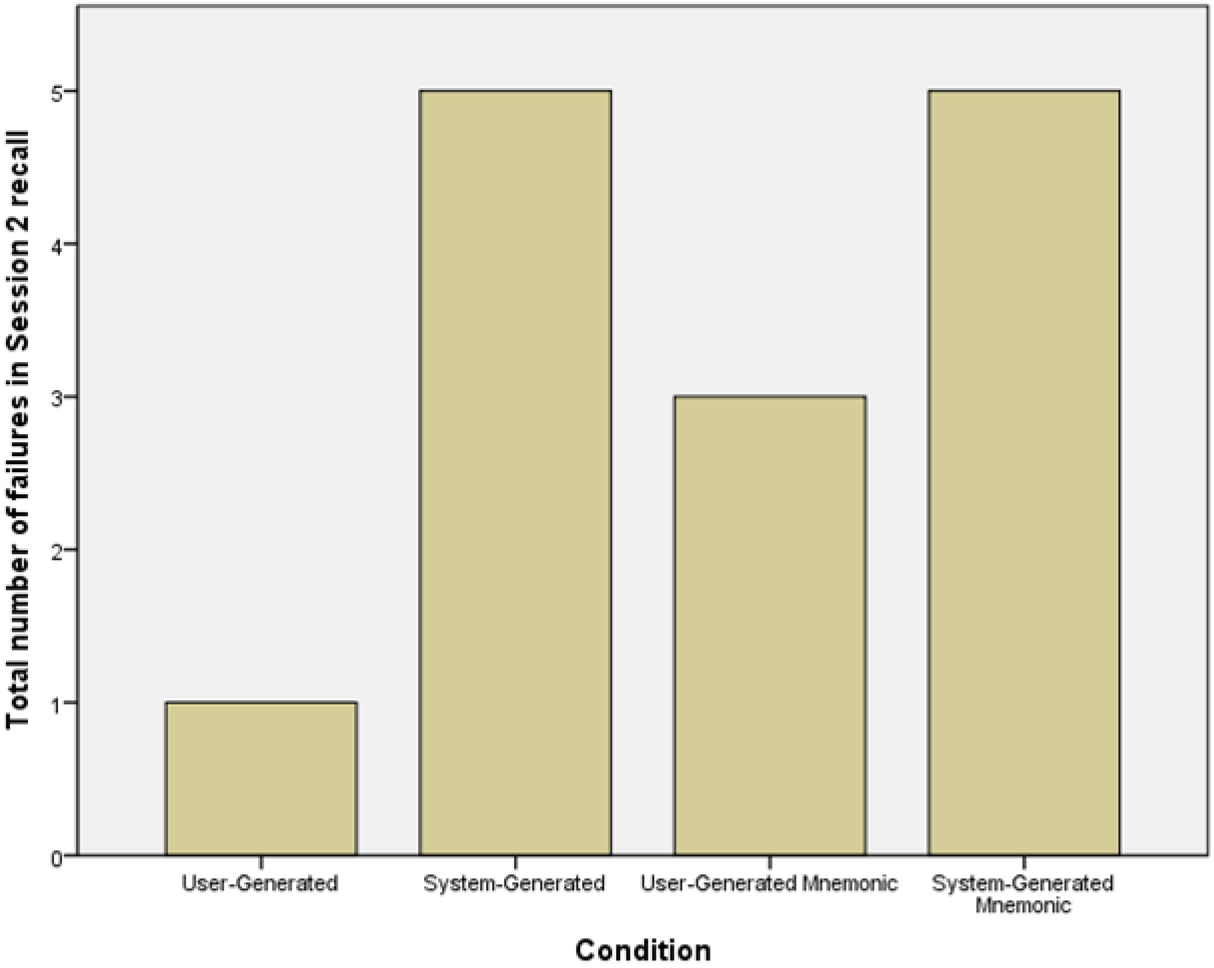


Figure 5.6: Distribution of the participants failing in session 2

Recall Edit Distance

The recall edit distance is measured using two dependent measures, the Damerau-Levenstein edit distance and the Jaro-Winkler proximity edit distance. For all successful logins into the password application, the Damerau-Levenstein and Jaro-Winkler edit distance are 0 and 1, respectively. The edit distances other than 0 and 1were recorded for the recall tasks in both sessions when participants failed to recall their passwords.

However, since participants were required to recall their passwords in the first session in order to participate in the second, those who failed to do so did not complete the study. Consequently, only edit distances for the second session were statistically analyzed.

Damerau-Levenshtein edit distances

The Damerau-Levenshtein edit distance between the recalled and the stored passwords is the minimum number of operations (insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters) needed to transform recalled passwords into those stored. For the passwords incorrectly recalled in the 2nd session, one user-generated recorded value was 1; five system-generated recorded values were 3, 3, 1, 1 and 5; three user-generated mnemonic policy values were 2, 3 and 2; and the five system-generated mnemonic policy values were 1, 2, 1, 3 and 5. The remaining passwords that were correctly recalled recorded a value of zero. Figure 5.7 shows the mean Damerau-Levenshtein distance distribution of passwords recalled in the Session 2 across the password policies.

Data for this dependent variable were non-normal. After reciprocal transformation, the skewness value remained lower than -2 with a high kurtosis value. These data suggest that this dependent variable was zero inflated as seventy-eight percent of the data had a value of zero.

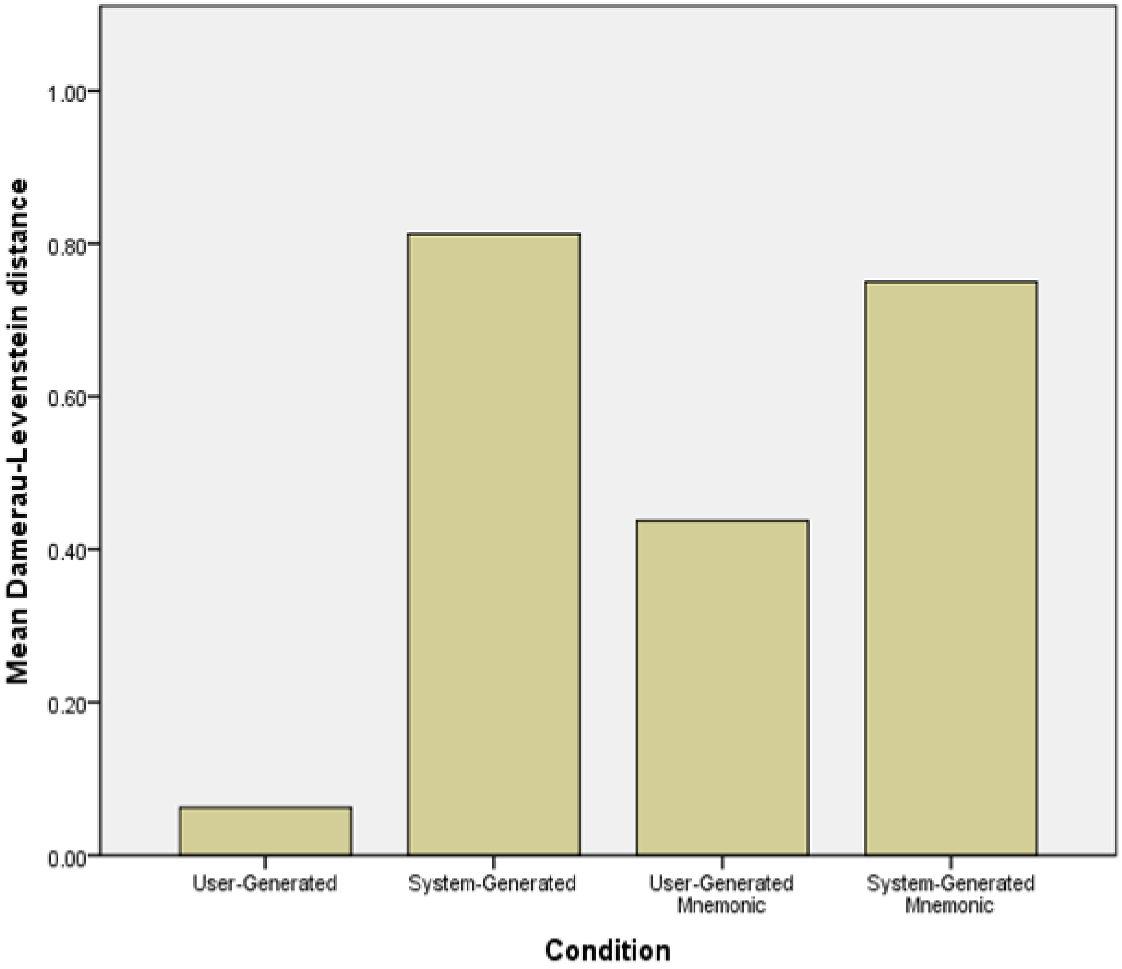


Figure 5.7: Distribution of mean Damerau-Levenstein edit distance of Session 2 recall passwords

Jaro-Winkler Proximity

The Jaro-Winkler proximity is a measure of the difference between the stored and the recalled passwords. From the passwords incorrectly recalled in the 2nd session, one user-generated policy recorded value was 0.967; five system-generated policy recorded values were 0.81, 0.746, 0.905, 0.905, and 0.631; three user-generated mnemonic policy recorded values were 0.849, 0.783 and 0.952; and the five system-generated mnemonic policy recorded of values were 0.897, 0.743, 0.905, 0.905, and 0.508. The remaining passwords that were recalled correctly were assigned a value of one. Figure 5.8 shows the mean Jaro-Winkler proximity edit distance distribution of the passwords recalled in the second session across the various password policies.

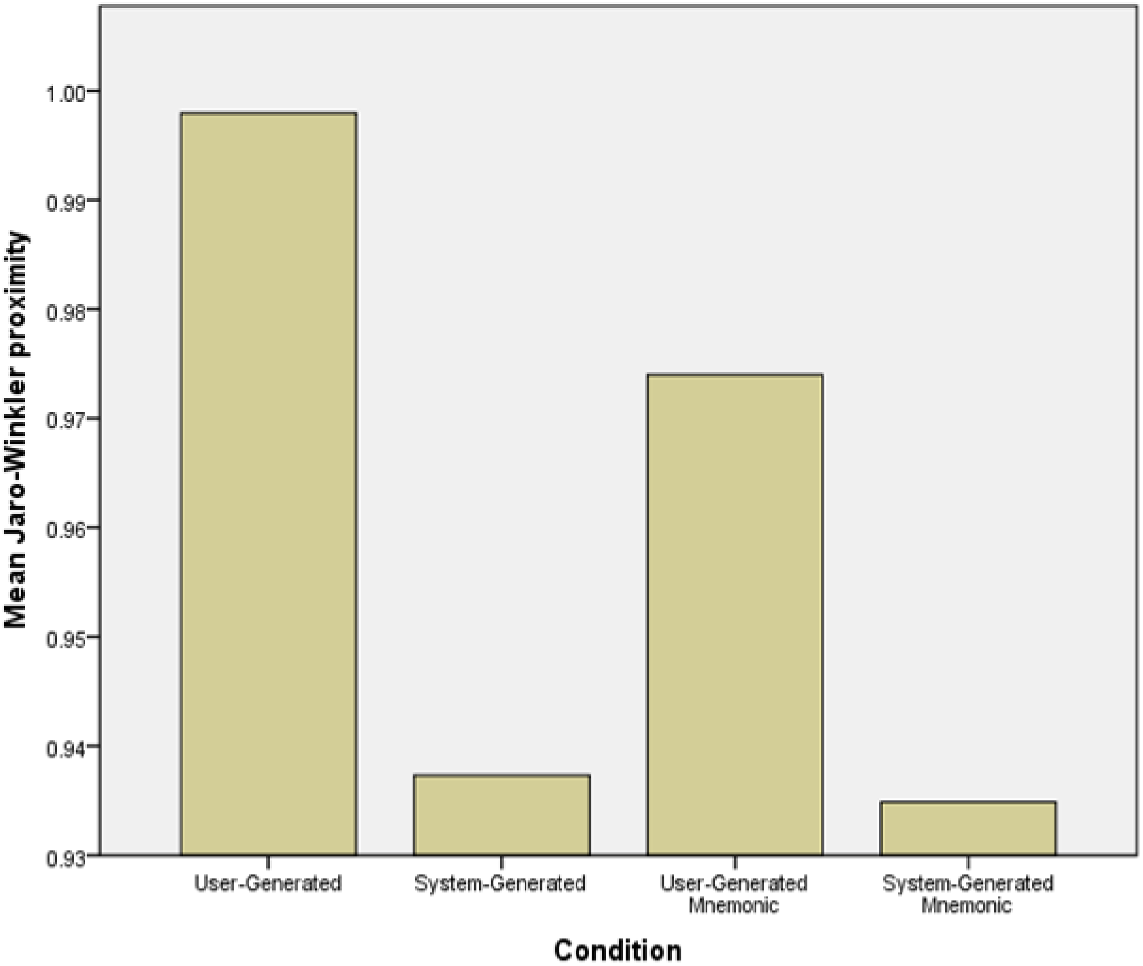


Figure 5.8: Distribution of mean Jaro-Winkler edit distance of Session 2 recall passwords

Data for this dependent variable were also non-normal. After reciprocal transformation, the skewness value remained higher than +2 along with a high kurtosis value. These data suggest that this dependent variable was one inflated, with seventy-eight percent of the data having a value of one.

Subjective Measures

The subjective measures recorded in the first session were also divided into two parts, one being the creation/memorization task and the other the recall of the password after a 5 minute interval. Subjective data were collected from the participants by recording their responses to the NASA TLX questionnaire and the SUS questionnaire for the creation/memorization task. Similarly, NASA TLX and SUS questionnaire scores were collected for the recall task. In the second session where the participants recalled their passwords after a week's time, the recall task NASA TLX questionnaire and SUS questionnaire scores were recorded.

Creation/Memorization task NASA TLX

This subjective measure was used to measure the task workload on the participants on the six 7-point scales of mental demand, physical demand, temporal demand, performance, effort and frustration. The description of each subscale is provided in Table 5.16.

The analysis of the data collected found that all NASA TLX measures were normally distributed. Then, a one-way ANOVA was conducted for each of those parameters, the results finding no significant differences among the password policies with respect to mental demand, physical demand, temporal demand, effort and frustration. In the case of performance, the analysis showed that there was a significant difference among password policies (F (3,63)=3.608,p=0.027). An LSD post-hoc test revealed that participants in the user-generated and system-generated password conditions felt they performed better than participants using the user-generated mnemonic and system-generated mnemonic password policies as shown in Table 5.17 and Figure 5.9.

Table 5.15: NASA-TLX rating scale definitions (Hart, 2002)

Title	Endpoints	Descriptions
Mental Demand	Low/High	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	Low/High	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low/High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Table 5.16 : LSD Post-hoc test on creation/memorization NASA TLX performance metric

(I) Condition	(J) Condition	Mean Difference (I-J)	Std. Error	Sig.
User-Generated	System-Generated	.00000	.51184	1.000
	User-Generated Mnemonic	-1.12500*	.51184	.032
	System-Generated Mnemonic	-1.25000*	.51184	.018
System-Generated	User-Generated	.00000	.51184	1.000
	User-Generated Mnemonic	-1.12500*	.51184	.032
	System-Generated Mnemonic	-1.25000*	.51184	.018
User-Generated Mnemonic	User-Generated	1.12500*	.51184	.032
	System-Generated	1.12500*	.51184	.032
	System-Generated Mnemonic	-.12500	.51184	.808
System-Generated Mnemonic	User-Generated	1.25000*	.51184	.018
	System-Generated	1.25000*	.51184	.018
	User-Generated Mnemonic	.12500	.51184	.808

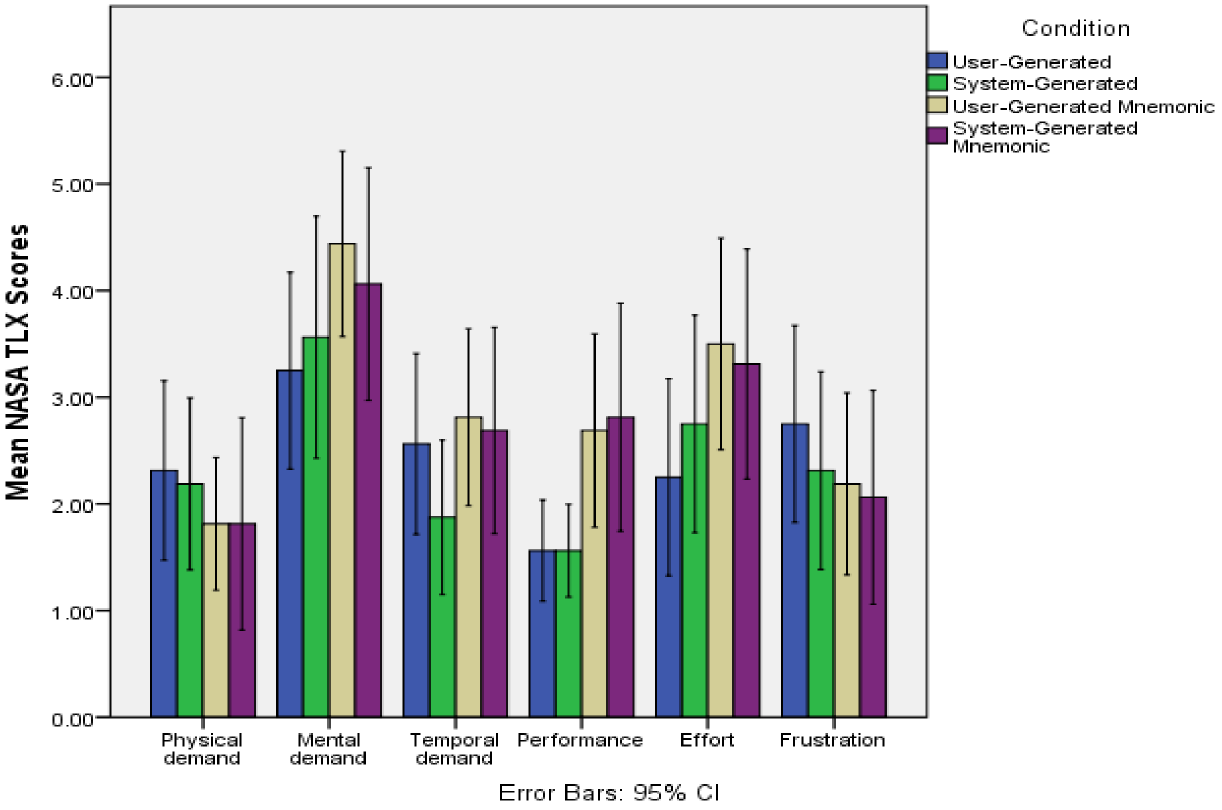


Figure 5.9: Mean NASA TLX measures for creation / memorization task

Creation/Memorization SUS

This subjective measure was used to determine the overall system usability by calculating a total usability score out of 100 from the responses given by the participants for the 10 questions after the creation/memorization task in the first session. The data collected were then analyzed for normality, the results indicating they were normal. The descriptive statistics of the SUS scores for the password creation/memorization task are provided in Table 5.18. Then, one-way ANOVA was used to check for a significant effect of password policy. Table 5.19 and Figure 5.10 below show no significant effect (F

(3, 63) = 1.850, p = 0.148) of password policy on the usability of the password creation task.

Table 5.17: Descriptive statistics of the SUS scores for password creation/memorization task

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	74.0625	20.61300	5.15325
System-Generated	16	58.2813	22.63329	5.65832
User-Generated Mnemonic	16	65.3125	16.50442	4.12610
System-Generated Mnemonic	16	70.0000	19.45079	4.86270
Total	64	66.9141	20.32349	2.54044

Table 5.18: One-way ANOVA of the SUS scores for password creation/memorization task

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2203.418	3	734.473	1.850	.148
Within Groups	23818.359	60	396.973		
Total	26021.777	63			

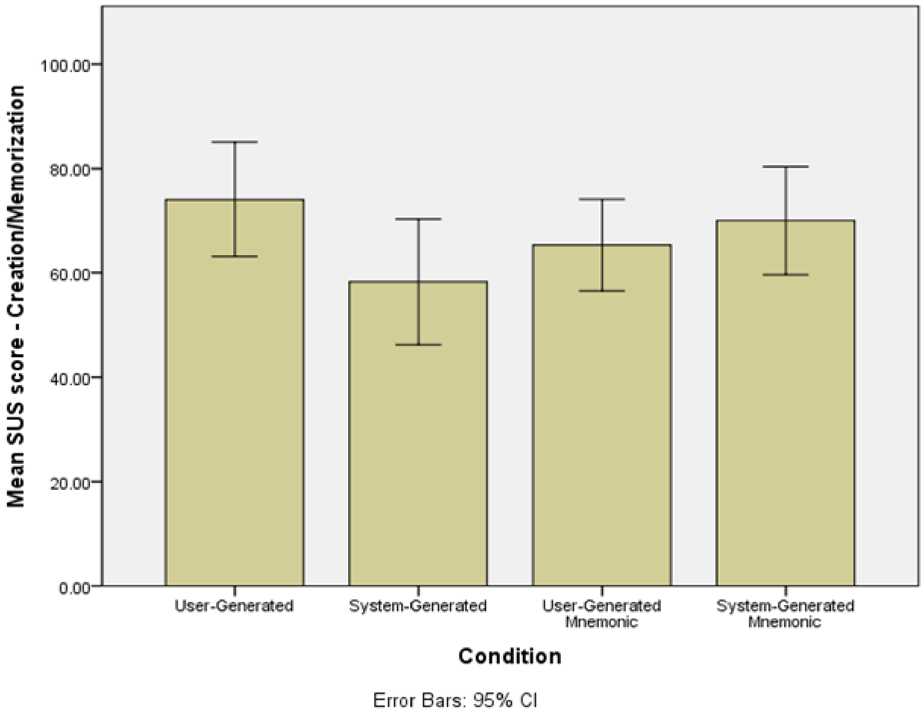


Figure 5.10: Mean SUS for creation / memorization task

Recall task NASA TLX

The NASA TLX assesses workload on the six 7-point scales of mental, physical and temporal loads, performance, effort, and frustration with low and high end points. The NASA TLX questionnaires were administered at the end of each recall task session, i.e., after the 1st session--recall and 2nd session--recall.

***Mental Demand*:** A two-way mixed ANOVA was conducted to test the main and interaction effects of password policy and task session on the mental demand experienced by the participants while recalling passwords. The results indicated the main effect of task session was significant, F (1, 60)=5.298, *p*=0.025, but the main effect of the password policy was not significant, F (3, 60)=1.240, *p*>0.05. Subsequent post-hoc analysis of the within-subject main effects revealed that mental demand was higher for recall in the second session than for recall in the first session (*p*=0.025) as shown in Figure 5.11. The interaction effect was not significant, F(3, 60)=0.582, *p*>0.05. The descriptive statistics and two-way ANOVA data for mental demand are provided in Tables 5.20, 5.21 and 5.22:

Table 5.19: Descriptive statistics for mental demand during recall in 1st session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	2.3125	1.66208	.41552
System-Generated	16	3.5000	1.93218	.48305
User-Generated Mnemonic	16	2.9375	1.76895	.44224
System-Generated Mnemonic	16	3.2500	1.98326	.49582
Total	64	3.0000	1.85164	.23146

Table 5.20: Descriptive statistics for mental demand during recall in 2nd session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	3.1250	1.89297	.47324
System-Generated	16	3.8750	2.57876	.64469
User-Generated Mnemonic	16	3.3750	2.30579	.57645
System-Generated Mnemonic	16	4.2500	2.38048	.59512
Total	64	3.6563	2.29021	.28628

Table 5.21: Two-way ANOVA data for mental demand

Mental Demand	SS	df	Mean Squares	F	Sig.
Task Sessions	13.781	1	13.781	5.298	0.025
Conditions	22.656	3	7.552	1.240	0.303
Task Sessions x Conditions	2.156	3	0.719	0.276	0.842
Error (Within-subject)	156.062	60	2.601		
Error (Between-subject)	365.562	60	6.093		

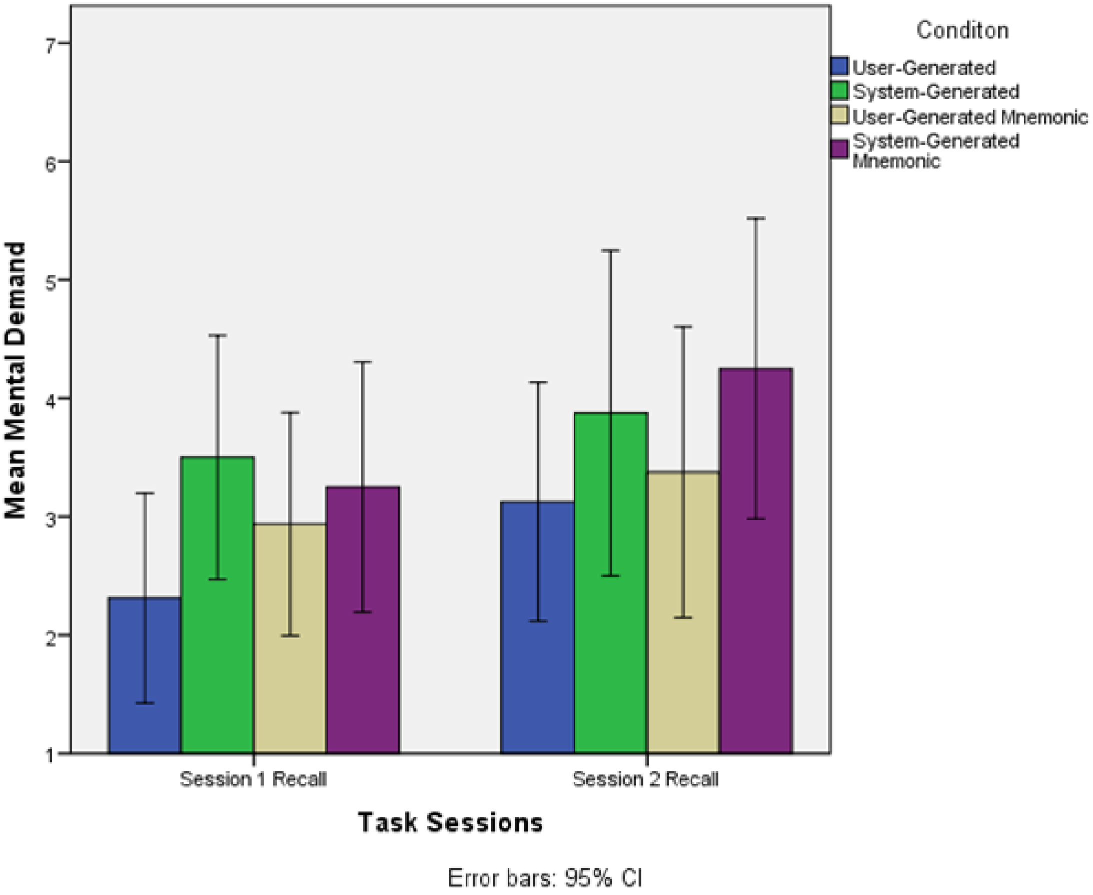


Figure 5.11: Mean rating for mental demand

***Physical Demand*:** A two-way mixed ANOVA was conducted to test the main and interaction effects of password policy condition and task session on the physical demand experienced by participants while recalling passwords. The results indicated the main effects were not significant, F(1,60)=0.621, *p*=0.434 for task sessions and F(3,60)=0.915, *p*=0.439 for password policies, as shown in Figure 5.12. The interaction effect was not significant, F(3,60)=0.080, *p*=0.970. The descriptive statistics and two-way ANOVA data for physical demand are provided in Tables 5.23, 5.24 and 5.25:

Table 5.22: Descriptive statistics for physical demand during recall in 1st session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	1.6875	1.35247	.33812
System-Generated	16	1.8750	1.45488	.36372
User-Generated Mnemonic	16	1.5000	1.03280	.25820
System-Generated Mnemonic	16	1.3125	.87321	.21830
Total	64	1.5938	1.19149	.14894

Table 5.23: Descriptive statistics for physical demand during recall in 2nd session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	1.8125	1.27639	.31910
System-Generated	16	2.0000	1.26491	.31623
User-Generated Mnemonic	16	1.5000	.89443	.22361
System-Generated Mnemonic	16	1.5000	.89443	.22361
Total	64	1.7031	1.09370	.13671

Table 5.24: Two-way ANOVA data for physical demand

Physical Demand	SS	df	Mean Squares	F	Sig.
Task Sessions	0.383	1	0.383	0.621	0.434
Conditions	22.656	3	1.862	0.915	0.439
Task Sessions x Conditions	0.148	3	0.049	0.080	0.970
Error (Within-subject)	36.969	60	0.616		
Error (Between-subject)	122.094	60	2.035		

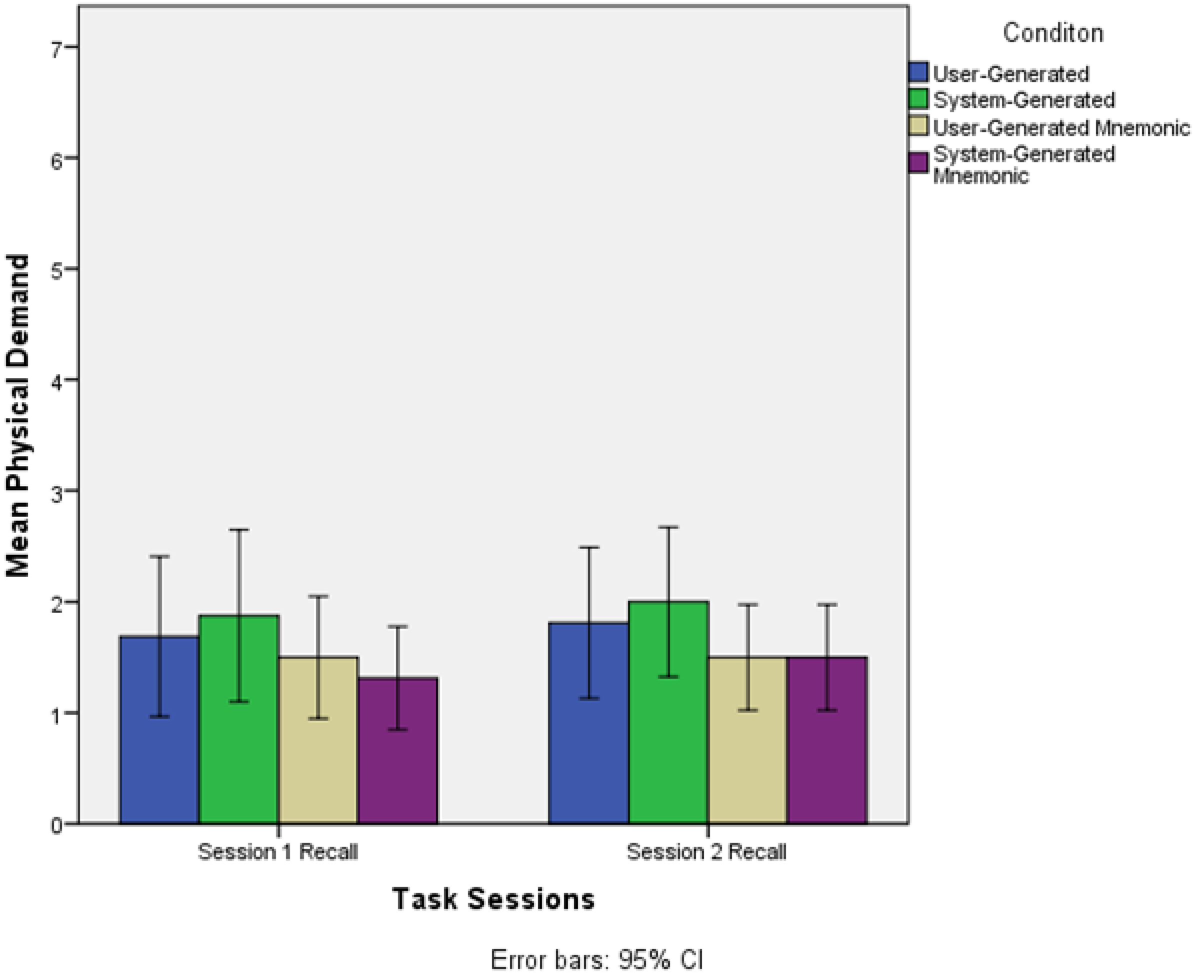


Figure 5.12: Mean rating for physical demand

Temporal Demand: A two-way mixed ANOVA was conducted to test the main and interaction effects of the password policy condition and task session on the temporal demand experienced by participants while recalling passwords. The results indicated the main effects were not significant, F(1,60)=0.090, *p*=0.766 for task sessions and F(3,60)=0.347, *p*=0.792 for password conditions as shown in Figure 5.13. The interaction effect was not significant, F(3,60)=0.595, *p*=0.621. The descriptive statistics and two-way ANOVA data for temporal demand are provided in Tables 5.26, 5.27 and 5.28:

Table 5.25: Descriptive statistics for temporal demand during recall in 1st session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	2.0000	1.46059	.36515
System-Generated	16	1.6875	1.35247	.33812
User-Generated Mnemonic	16	1.9375	1.23659	.30915
System-Generated Mnemonic	16	1.5625	.81394	.20349
Total	64	1.7969	1.22383	.15298

Table 5.26: Descriptive statistics for temporal demand during recall in 2nd session

	N	Mean	Std. Deviation	Std. Error
User Generated	16	1.6875	1.19548	.29887
System Generated	16	1.8125	1.55858	.38964
User Generated Mnemonic	16	2.1250	1.31022	.32755
System Generated Mnemonic	16	1.7500	.93095	.23274
Total	64	1.8438	1.25000	.15625

Table 5.27: Two-way ANOVA data for temporal demand

Temporal Demand	SS	df	Mean Squares	F	Sig.
Task Sessions	0.070	1	0.070	0.090	0.766
Conditions	2.461	3	0.820	0.347	0.792
Task Sessions x Conditions	1.398	3	0.466	0.595	0.621
Error (Within-subject)	47.031	60	0.784		
Error (Between-subject)	141.906	60	2.365		

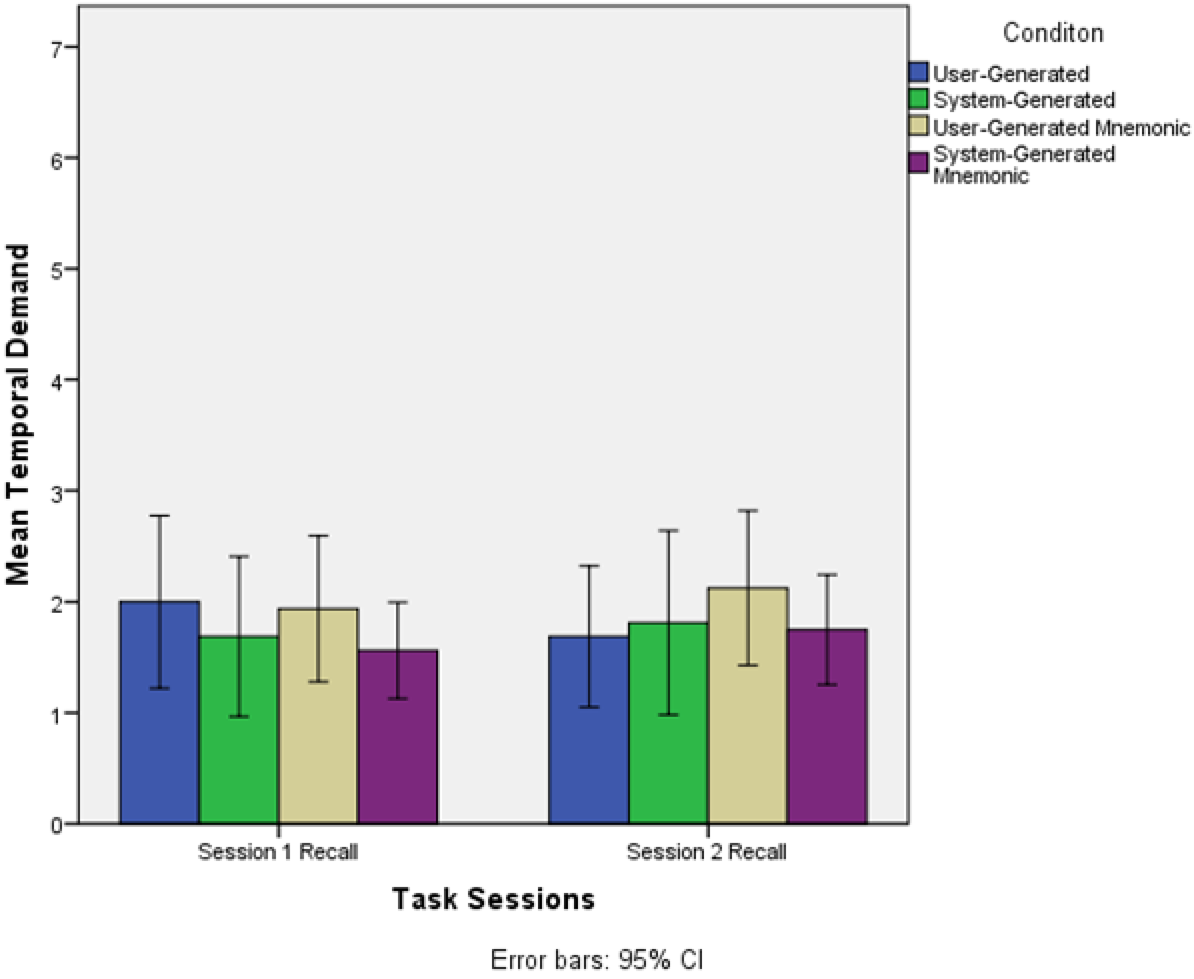


Figure 5.13: Mean rating for temporal demand

Performance: A two-way mixed ANOVA was conducted to test the main and interaction effects of password policy and task session on the performance component of the NASA-TLX while recalling passwords. The results indicated the main effect of the task session was significant, F(1,60)=8.216, *p*=0.006 and main effect of the password policy was not significant, F(3,60)=1.297, *p*=0.284. The performance component was higher for recall in the second session than for recall in the first session (*p*=0.006) as shown in Figure 5.14, indicating that participants were less satisfied with their performance in the second session. The interaction effect was not significant, F(3,60)=1.228, *p*=0.308. The descriptive statistics and two-way ANOVA data for performance are provided in Tables 5.29, 5.30 and 5.31:

Table 5.28: Descriptive statistics for performance during recall in 1st session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	1.3125	.79320	.19830
System-Generated	16	1.0625	.25000	.06250
User-Generated Mnemonic	16	1.8125	1.64190	.41047
System-Generated Mnemonic	16	1.6250	1.45488	.36372
Total	64	1.4531	1.18093	.14762

Table 5.29: Descriptive statistics for performance during recall in 2nd session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	1.5000	1.09545	.27386
System-Generated	16	2.6875	2.67628	.66907
User-Generated Mnemonic	16	2.2500	2.40832	.60208
System-Generated Mnemonic	16	3.0000	2.70801	.67700
Total	64	2.3594	2.33243	.29155

Table 5.30: Two-way ANOVA data for performance

Performance	SS	df	Mean Squares	F	Sig.
Task Sessions	26.281	1	26.281	8.216	0.006
Conditions	13.812	3	4.604	1.297	0.284
Task Sessions x Conditions	11.781	3	3.927	1.228	0.308
Error (Within-subject)	191.937	60	3.199		
Error (Between-subject)	213.062	60	3.551		

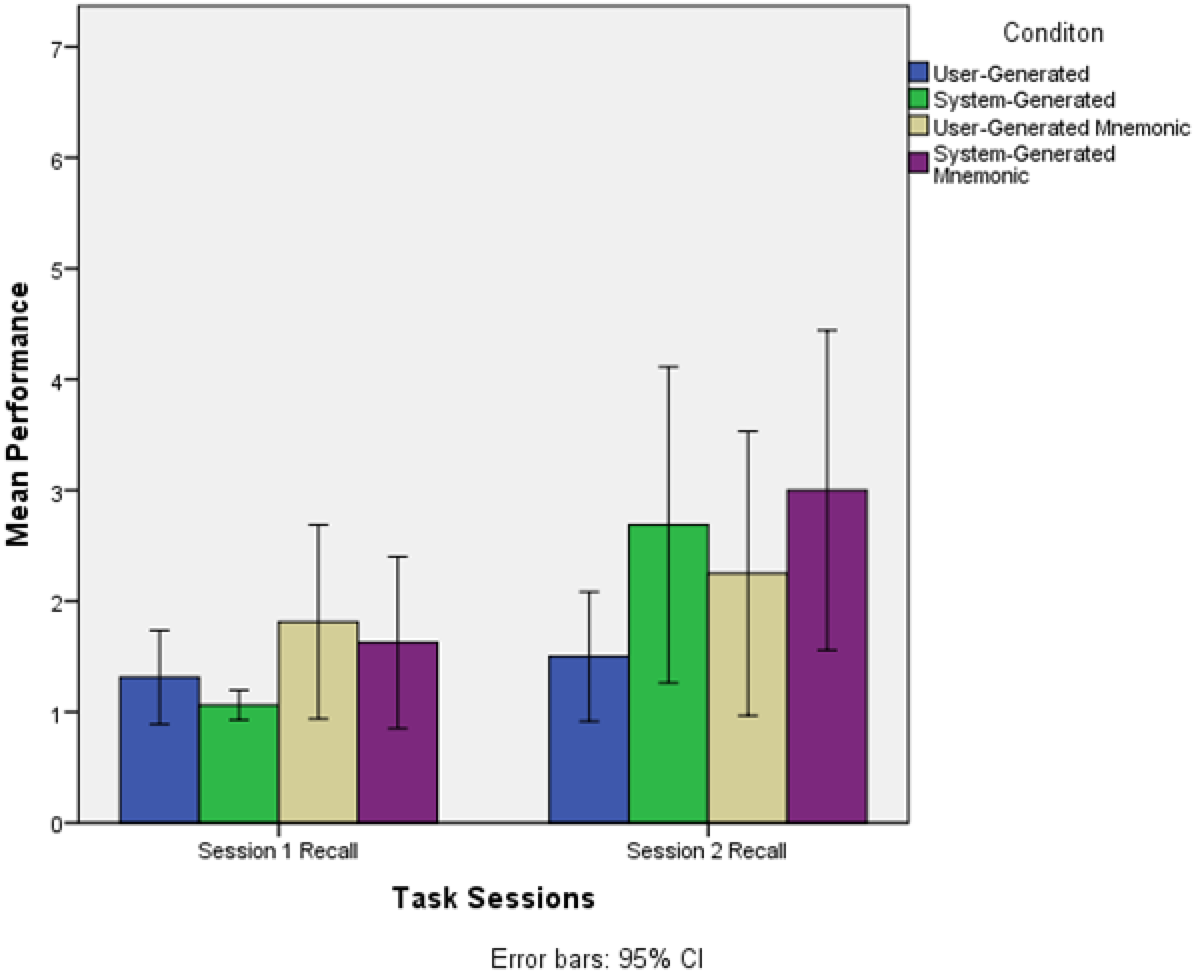


Figure 5.14: Mean rating for performance

Effort: A two-way mixed ANOVA was conducted to test the main and interaction effects of the password policy condition and task session on the effort experienced by participants while recalling passwords. The results indicated main effects were not significant, F(1,60)=3.549, *p*=0.064 for task sessions and F(3,60)=0.593, *p*=0.622 for password conditions as shown in Figure 5.15. The interaction effect was not significant, F(3,60) =0.184, *p*=0.907. The descriptive statistics and two-way ANOVA data for effort are provided in Tables 5.32, 5.33 and 5.34:

Table 5.31: Descriptive statistics for effort during recall in 1st session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	2.0000	1.59164	.39791
System-Generated	16	2.1250	1.78419	.44605
User-Generated Mnemonic	16	2.6875	1.66208	.41552
System-Generated Mnemonic	16	2.5625	1.75000	.43750
Total	64	2.3438	1.68296	.21037

Table 5.32: Descriptive statistics for effort during recall in 2nd session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	2.3750	1.54380	.38595
System-Generated	16	2.8750	2.15639	.53910
User-Generated Mnemonic	16	2.9375	2.01556	.50389
System-Generated Mnemonic	16	3.0625	2.14379	.53595
Total	64	2.8125	1.95078	.24385

Table 5.33: Two-way ANOVA data for effort

Effort	SS	df	Mean Squares	F	Sig.
Task Sessions	7.031	1	7.031	3.549	0.064
Conditions	8.594	3	2.865	0.593	0.622
Task Sessions x Conditions	1.094	3	0.365	0.184	0.907
Error (Within-subject)	118.875	60	1.981		
Error (Between-subject)	289.625	60	4.827		

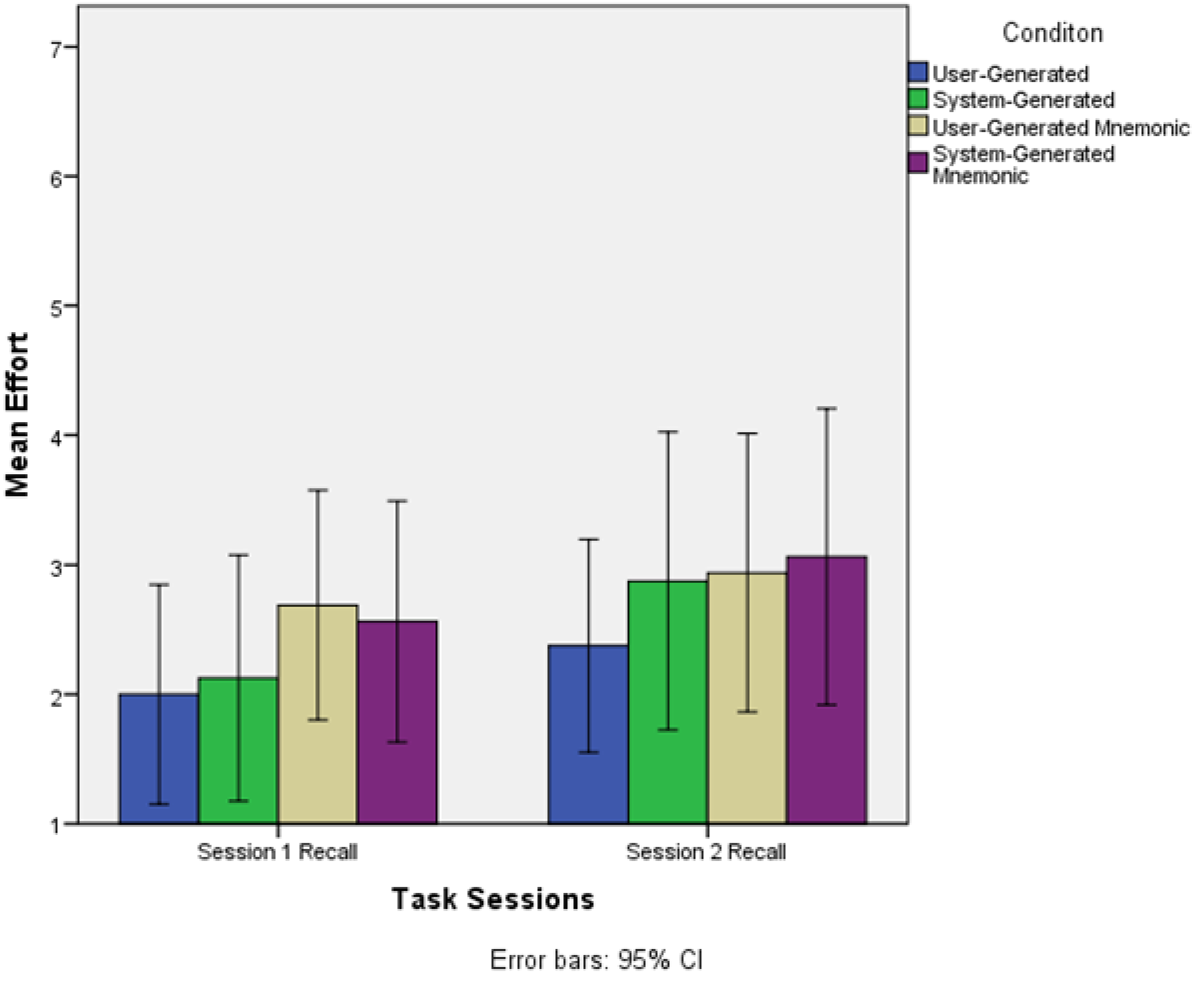


Figure 5.15: Mean rating for effort

Frustration: A two-way mixed ANOVA was conducted to test the main and interaction effects of password policy and task session on the frustration component of the NASA-TLX while recalling passwords. The results indicated the main effect of task session was significant, F(1, 60)=4.021, *p*=0.049, but the main effect of the password policy was not significant, F(3, 60)=0.338, *p*=0.798. The frustration component was higher for recall in the second session than for recall in the first session (*p*=0.049) as shown in Figure 5.16. The interaction effect was not significant, F(3, 60)=0.991, *p*=0.403. The descriptive statistics and two-way ANOVA data for performance are provided in Tables 5.35, 5.36 and 5.37:

Table 5.34: Descriptive statistics for frustration during recall in 1st session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	1.9375	1.34009	.33502
System-Generated	16	2.1875	1.32759	.33190
User-Generated Mnemonic	16	1.8750	1.14746	.28687
System-Generated Mnemonic	16	1.9375	1.69189	.42297
Total	64	1.9844	1.36268	.17033

Table 5.35: Descriptive statistics for frustration during recall in 2nd session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	2.1250	1.45488	.36372
System-Generated	16	2.5000	2.03306	.50827
User-Generated Mnemonic	16	2.8750	1.92787	.48197
System-Generated Mnemonic	16	2.0625	1.61116	.40279
Total	64	2.3906	1.76039	.22005

Table 5.36: Two-way ANOVA data for frustration

Frustration	SS	df	Mean Squares	F	Sig.
Task Sessions	5.281	1	5.281	4.021	0.049
Conditions	3.812	3	1.271	0.338	0.798
Task Sessions x Conditions	3.906	3	1.302	0.991	0.403
Error (Within-subject)	78.812	60	1.314		
Error (Between-subject)	225.687	60	3.761		

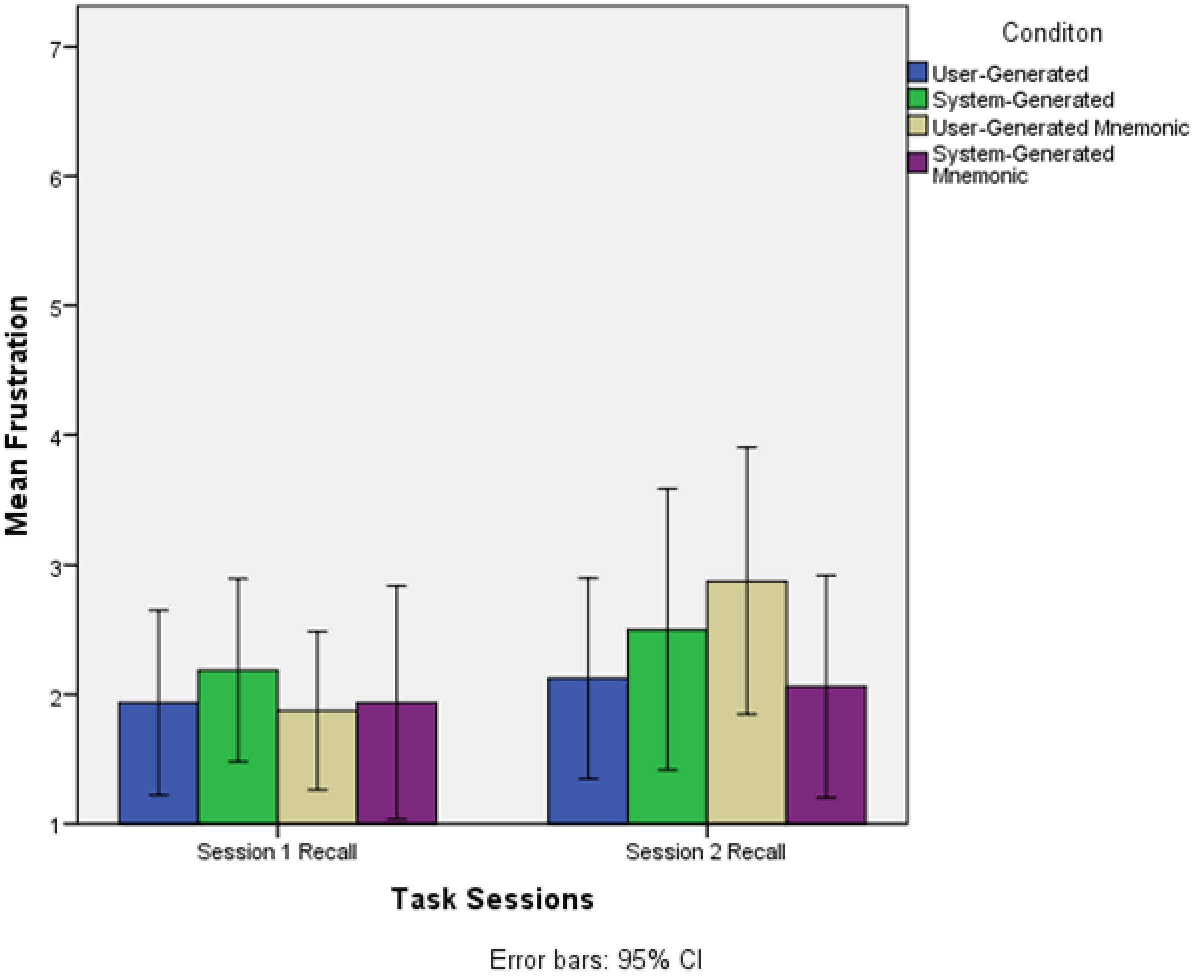


Figure 5.16: Mean rating for frustration

Recall task SUS

The SUS questionnaires were administered at the end of each recall task, i.e., 1st session--recall and 2nd session--recall. The descriptive statistics for the SUS scores for the password recall task for each session are provided in Tables 5.38 and 5.39:

Table 5.37: Descriptive statistics of the SUS scores for password recall task in 1st session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	74.6875	19.74578	4.93645
System-Generated	16	57.0313	24.20776	6.05194
User-Generated Mnemonic	16	68.5938	19.74776	4.93694
System-Generated Mnemonic	16	70.3125	20.38944	5.09736
Total	64	67.6563	21.62026	2.70253

Table 5.38: Descriptive statistics of the SUS scores for password recall task in 2nd session

	N	Mean	Std. Deviation	Std. Error
User-Generated	16	69.2188	21.40325	5.35081
System-Generated	16	55.6250	23.81351	5.95338
User-Generated Mnemonic	16	62.5000	22.24860	5.56215
System-Generated Mnemonic	16	64.2188	24.02635	6.00659
Total	64	62.8906	22.87916	2.85990

A two-way mixed ANOVA was conducted to test the main and interaction effects of the password policy conditions and the task sessions on system usability while recalling passwords. The results indicated that the main effect of task session was significant, F(1, 60)= 5.214, *p*=0.026, but the main effect of password policy was not significant, F(3, 60)=1.653, *p*=0.187. Post-hoc analysis of the task session main effect revealed that the

SUS score was higher during the first recall session than during the recall of the same password in the second session (*p*=0.026). Figures 5.17 and 5.18 present the mean SUS scores for the password policy conditions for the first and second sessions, respectively. The interaction effect was not significant, F(3, 60)=0.293, *p*=0.830. The two-way ANOVA data for the SUS scores are provided in Table 5.40:

Table 5.39: Two-way ANOVA data for SUS score

SUS	SS	df	Mean Squares	F	Sig.
Task Sessions	726.758	1	726.758	5.214	0.026
Conditions	4117.773	3	1372.591	1.653	0.187
Task Sessions x Conditions	122.461	3	40.820	0.293	0.830
Error (Within-subject)	8363.281	60	139.388		
Error (Between-subject)	49822.656	60	830.378		

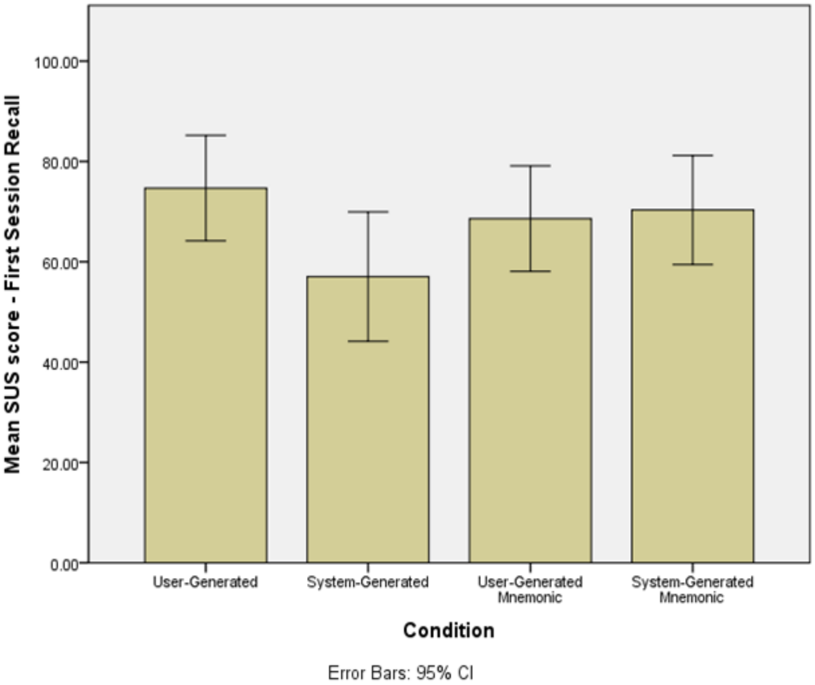


Figure 5.17: Mean SUS for first session recall

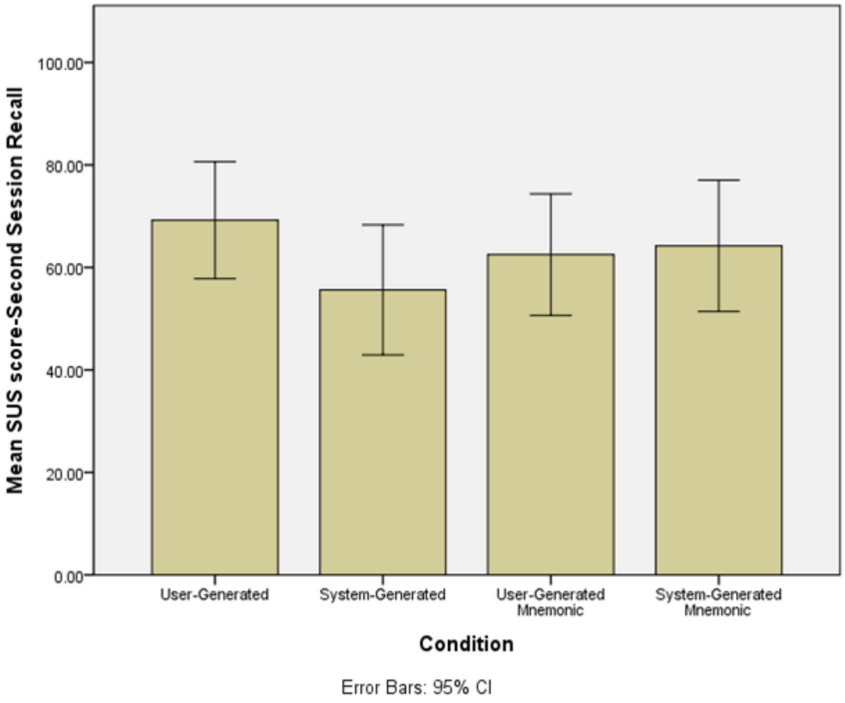


Figure 5.18: Mean SUS for second session recall

Power Analysis

G*Power software (Erdfelder, Buchner, Lang, 2009) was used to conduct a power analysis to calculate the sample size required to produce significance among conditions. All the dependent measures were tested for required sample size to obtain significance except the ones which already had significant differences, such as creation and/or memorization time and creation NASA TLX performance measure. The least number of samples required to obtain a significant difference was 180 total participants as shown in the Figure 5.19 below.

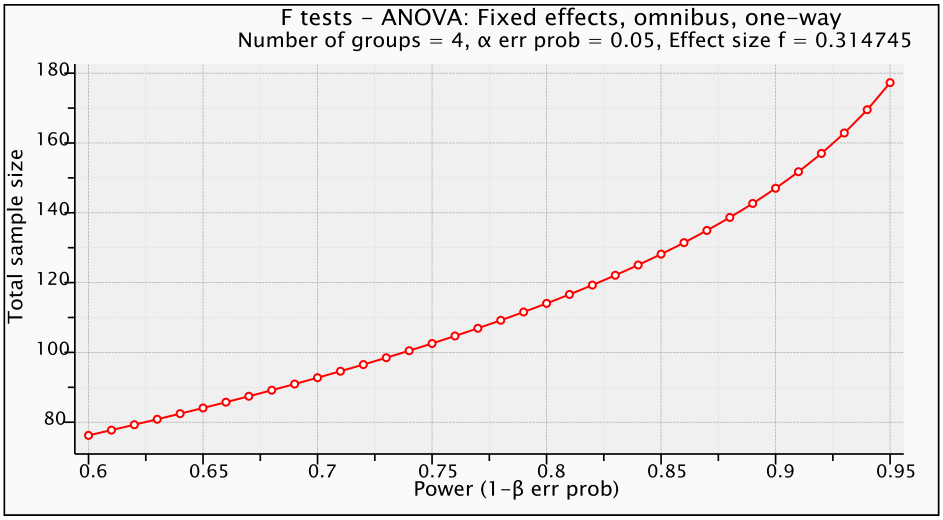


Figure 5.19: Power analysis of Session 1 recall time

6. DISCUSSION

The main objective of this research was to compare the usability of a novel system-generated mnemonic policy with three existing policies, while maintaining a constant level of security across the policies. The usability of these policies was measured across three tasks, password creation and/or memorization, password recall after 5 minutes and password recall after a week. Dependent measures with respect to the tasks were collected and statistically analyzed as shown in the results section.

In this study, to track the ease of creating a password, the dependent measures included the creation and/or memorization time, error rate, SUS, and NASA TLX. The memorability of the passwords was determined using the dependent measures of recall time, error rate, SUS, edit distance (Damerau-Levenshtein edit distance and Jaro-Winkler proximity) and NASA TLX for both the recall of the password after 5 minutes and after a week.

The statistical analysis of the collected data as shown in the results section demonstrates a significant difference between the password policies for the creation and/or memorization time of the password and for the creation performance metric in the NASA TLX dependent measure. To identify potential explanations for these results, comments from the participants and personal observations of the facilitator were used. The results for each task session are discussed in the following sections.

Creation and/or memorization task:

Among all the password policies for the creation and/or memorization time metric, the user-generated mnemonic policy participants took significantly more time (207.13 seconds) to memorize their password than any other participants. This is because creating their own mnemonic based on the training provided was a mentally demanding and time-consuming task. There were no significance differences among the other three policies: the system-generated mnemonic policy had the next highest mean creation and/or memorization time of 70.52 seconds followed by the user-generated password (61.97 seconds) and system-generated password (54.43 seconds) policies. This finding suggests that providing no mnemonic aid to participants results in less time taken in creating and/or memorizing passwords. Based on this creation and/or memorization time metric, the most efficient method for creating a password is the system-generated password policy: assign the password to the users and ask them to memorize it using their own techniques without providing any aid.

The creation error rate measures how effectively participants create passwords without errors. There was no significant difference among the password policies for this metric. The system-generated password policy had the highest mean error rate (0.10) followed by the user-generated password (0.06), user-generated mnemonic (0.03), and system-generated mnemonic (0) policies. The application displayed the password assigned to the participants while they created their account. This may have helped them to create their account without errors, resulting in low overall error rates. The work load of creation and/or memorization of the password was measured using the NASA TLX questionnaire at the end of this task. Performance was the only NASA TLX metric to show a significant difference. The subsequent post-hoc analysis revealed that the participants in the user-generated password and system-generated password policies believed that they performed significantly better than the participants in the mnemonic-based policies.

Based on the SUS, the usability of the creation and/or memorization task did not differ significantly across the policies, with the results showing that the highest mean SUS score was for the user-generated password policy (74.06) followed by the system-generated mnemonic (70.00), the user-generated mnemonic (65.32), and the system-generated password (58.28) policies. One of the reasons for this finding could be that the participants were already familiar with the user-generated password policy as it is the most commonly used, and, therefore, they found it easy-to-use. Because the system-generated mnemonic technique provided users with assistance for remembering their password, they may have believed it to be more usable than the user-generated mnemonic technique. Because the system-generated password policy was composed of random letters and did not provide any memory aid, users may have believed it was the least usable. This finding is partially supported by Zviran and Haga (1993) who found that user-generated passwords were more usable than assigned system-generated passwords.

The overall usability level of the creation and/or memorization of passwords using the proposed system-generated mnemonic policy were neither significantly better nor worse than any other policy. This policy does not take significantly less time to create a password, nor does it have a significantly lower error rate, workload, or SUS score during the password creation phase. Thus, there is no benefit during this phase in using this policy. Therefore this policy cannot be recommended over the simpler and commonly used system-generated and user-generated password policies on the basis of password creation performance.

First session recall task:

The short-term memorability of the passwords was the focus of the first session recall task. It was measured by the number of participants in each condition failing to recollect their passwords after playing Angry Birds™ (Lehtinen, 2009) for 5 minutes. The resemblance of the incorrectly recollected password to the actual ones created in the previous creation task was measured using: the Damerau-Levenshtein edit distance and Jaro-Winkler proximity.

It was observed that both user-generated policies (user-generated password policy and user-generated mnemonic policy) performed better in the short-term memorability metric. The participants using the system-generated policies (system-generated password policy and system-generated mnemonic policy) had the highest failure rates for recollecting their password. 18.75% of participants in the system-generated password condition and 25% in the system generated mnemonic condition failed to recollect the password on the first attempt. Among the participants assigned the user-generated policies, only 6.25% failed to recollect their password on the first attempt. The difference in these percentages indicates that the system-generated passwords were less memorable than those created using the other two policies. This conclusion is supported by the responses to the exit survey. The demographic data revealed that the majority of the participants who failed to recall their password in the system-generated mnemonic policy condition were non-native English speakers. This may have been a contributing factor for their failure.

A qualitative analysis of the user comments on the exit survey revealed that several of the participants believed that the words used in the system-generated mnemonic were difficult for them to remember, a typical example being the words starting with the letter *x*—Xenops, Xenophobic and Ximenias. According to one participant, “It was difficult for me to try and remember a meaningless long sentence with an awkward combination of words!” In addition, some participants said that they should be given the freedom of requesting a new password and mnemonic if they were not satisfied with the one assigned to them. For example, one of the participants in the system-generated mnemonic condition received the password “pwamxcx” with a generated mnemonic of “Peter's wild armadillo mainly xeroxed countless ximenias”. He commented, “The reason I couldn’t remember the mnemonic and password was because it was too awkward, confusing and meaningless to me.”

The average Damerau-Levenshtein edit distance was 2.50 for the system-generated mnemonic policy and 2.33 for the system-generated password policy. These values were more than twice as high as the two user-generated policies, which had an average value of 1. The average Jaro-Winkler proximity was 0.799 for the system-generated mnemonic policy and 0.778 for the system-generated password policy. The average values for the user-generated password and user-generated mnemonic conditions, were 0.893 and 0.905, respectively. These two metrics suggest that when participants failed to remember their password, they tended to be closer to being correct when using a user-generated password than when using system-generated password.

The remaining dependent variables for the first session recall task were recall time, error rate, NASA TLX, and SUS. None of these showed significant differences across password policies. As a whole, this analysis suggests that in terms of short-term memorability, the two policies that required the user to generate either a password or a mnemonic were more usable than the two policies in which either a password or a password and a mnemonic were assigned to the user.

Second session recall task:

The focus of the second session recall task was the long-term memorability of the passwords created and/or memorized using the password policies. It was measured using the same metrics used for the short-term memorability of the passwords.

Similar to the previous session recall results, both the system-generated policies (system-generated password policy and system-generated mnemonic policy) performed worse than the user-generated policies (user-generated password policy and user-generated mnemonic policy). However, in the second session, there were more failures overall than in the first session. Specifically, the participants in both of the system-generated policy groups had failure rates of 31.25%, while 18.75% of the user-generated mnemonic policy participants and 6.25% of the user-generated password policy participants failed in the second session recall. The exit survey found that 81.25% of the system-generated mnemonic participants believed that this method for creating passwords was awkward to use. Even though the system-generated mnemonic provided some meaning, it was difficult for the participants to relate to it personally. None of the system-generated password participants gave positive feedback on this policy. Seventyfive percent of the system-generated password participants said they used a chunking and pronunciation mnemonic technique to remember their assigned random password. Leonhard and Venkatakrishnan (2007) reported that among the random password generators they studied, the pronounceable password generator (PRONOUNCE3) was subjectively preferred, supporting this finding.

Similar to the results for the first recall session, the user-generated password policy participants were comfortable with the passwords they created and/or memorized. Only 6.25% of these participants failed to recall their password in the second session compared to 31.25% of the system-generated password group participants. The average Damerau-Levenshtein edit distance of the participants who failed to recall their password, was 2.40 for the system-generated mnemonic policy and 2.60 for the system-generated password policy. These values were more than twice those of the user-generated password policy, which had an average value of 1. However, the user-generated mnemonic policy had a value of 2.33, close to the value of the system-generated password policy. The average Jaro-Winkler proximity was 0.792 for the system-generated mnemonic policy and 0.799 for the system-generated password policy, while the average values of the user-generated password policy and the user-generated mnemonic policy were 0.967 and 0.861, respectively. The majority of the comments in the exit survey from the user-generated policy conditions suggested that since these participants created their own password and/ or mnemonic aid, they were able to remember them easily. The remaining dependent variables for the long-term recall task were recall time, error rate, NASA TLX and SUS. None of these showed significant differences across password policies. As a whole, neither of the system-generated policies is as usable as the user-generated policies in terms of the long-term memorability of the passwords.

Difference across task sessions:

The objective and subjective measures of the recall tasks for both sessions were analyzed with two-way ANOVA to examine the simple and interaction effects of the dependent variables with respect to password policies and sessions.

The dependent measures that showed significant differences between task sessions were recall time, error rate, the NASA TLX's mental demand and performance metrics, and SUS. The analysis showed that the recall time for the second session was significantly greater than for the first. Similarly, the second session error rate and the NASA TLX mental demand and performance metrics were significantly higher than for session one. The SUS score for session one was significantly higher than for session two. None of the dependent measures exhibited a significant interaction effect among task session and password policy conditions. These results suggest that participants performed worse in the second session than in the first session, presumably due to the degradation effect of time on memorability.

Analysis of user-generated passwords:

The results indicate that user-generated passwords appear to be more usable than system-generated passwords. One explanation for this may be that it is easier to remember a self-generated password than a randomly generated one. An analysis of the user-generated passwords in the study revealed that 93.75% of them contained words found in the dictionary or contained context that would be meaningful to others, as can be observed in Table 6.1:

Table 6.1: The user-generated passwords

Mindtree89!	1591964@Nl	techMahindra87$	Sarkar135$
Cedar@2010	Salmaka1!	Mega@10155	Samantha.E25
Leoroque30!	Greendude@7	Tacoma22@	Cl3mson@4
Angry$3578	Thavle123$%	!Clemson2011	Thimmaiah@10

In addition, 81.25% of them started with an upper case character at the beginning in order to satisfy the restriction that the password must contain an upper case character, and all the passwords had either a number or special character at the end to satisfy one of those restrictions. Participants apparently felt that it would be easier for them to remember a first-character capital letter and last character number or special character than it would be to remember these characters at some other position. While this practice allows for easy memorization, it has serious implications in terms of security because hackers might easily guess the position of the upper case or special character, thereby making these password restrictions less helpful in increasing security. Therefore, the user-generated policy restrictions should perhaps be modified to prevent such predictable user behaviors. But this, in turn, might dictate the usability of user-generated passwords.

7. CONCLUSION

In this study a new password policy was proposed in which a system generates a random password and an associated mnemonic. A computer application was built to generate a seven character random password and a mnemonic phrase from a list of predefined words. The main objective of this study was to evaluate the usability of this password policy and compare it to three existing password policies: user-generated password with restrictions, system-generated password with a user-generated mnemonic and system-generated password with no mnemonic. This research found that quantitatively the system-generated mnemonic policy was not statistically significantly different from the three other policies. However, the user-generated polices (user-generated password policy and user-generated mnemonic policy) tended to perform better than the system-generated policies (system-generated password policy and system-generated mnemonic policy).

The overall usability of the policies was measured using three user tasks: creation and/or memorization, short-term recall, and long-term recall. The system-generated mnemonic policy appeared to be as usable as the other policies for the creation/memorization task. However, in the recall tasks both user-generated policies performed better than either of the system-generated policies. Users tended to remember passwords or mnemonics that they created better than those assigned to them. The major disadvantage of creating one’s own passwords is that they tend to be predictable and thus less secure than randomly generated passwords. It was thought that a user-generated mnemonic policy in which the participants created their own mnemonic for a system-generated random password might enhance memorability while maintaining security. The most prevalent complaint regarding the user-generated mnemonic policy was that creating the mnemonic itself was a cognitively demanding task. However, these participants appeared to remember their password better than both system-generated password policies, in part perhaps simply because they spent more time memorizing it.

The exit survey found that 75% of the system-generated password policy participants used a chunking and pronouncing mnemonic technique to remember their password. This suggests that this is a common method people use to remember passwords. This chunking and pronouncing mnemonic technique could be utilized by password memory aid designers. Several participants in the system-generated mnemonic condition suggested the following design improvements:

- Provide a refresh button that assigns another password and mnemonic if the user is not comfortable with the one they have been assigned,
- Suggest a mnemonic and alternative word for each character of the password and give the user the control over choosing the words and constructing their own mnemonic. An example is shown in Figure 7.1 below.

<table>
<tr><td colspan="7">Password: ghcgsrp</td></tr>
<tr><td colspan="7">Mnemonic: George’s hungry cat gladly shared raw pancakes.</td></tr>
<tr><td>Garret’s
Glenn’s
Grant’s</td><td>Hilarious
Handsome
Happy</td><td>Cow
Camel
Cheetah</td><td>Gently
Gracefully
Grievingly</td><td>Showed
Shake
Shopped</td><td>Real
Ready
Rare</td><td>Pasta
Pastry
Peach</td></tr>
</table>

Figure 7.1: Mnemonic creation method suggested by users

In general, this research found that people tend to remember passwords and mnemonics they generated better than assigned ones. Even though the system-generated mnemonics were meaningful sentences, the participants could not relate as well to them as to passwords or mnemonics created themselves. However, the generalizability of the results of this study is limited by the following study constraints:

- More than 50% of the participants were non-native English speakers. Those participants might have experienced particular difficulty in memorizing the randomly generated mnemonics.
- The sample size (n = 16), of the study was small.
- No memory test screening was performed on the participants.

This research study is a first step in designing a system-generated password and mnemonic policy. Analyzing the qualitative and quantitative data from this study the following design suggestions are proposed, that could be followed while designing future system-generated mnemonic applications;

- Users must be given more control over the selection of a system-generated mnemonic for their assigned password if they are not satisfied with the one initially generated.
- Rather than providing only one mnemonic sentence, each character of the password could be given three to four word suggestions for the users to choose from to enable them to create the mnemonic best suited for them.
- The vocabulary created for the mnemonic generation application should be screened with potential users for their feedback before implementation in the system. In this way difficult words could be eliminated and the overall usability of the system-generated mnemonic increased.
- In order to fulfill NIST level 2 security standards for passwords, entropy of 30 bits or more has to be maintained. Therefore the original character set of 26 letters (32.9 bit entropy) in English could be reduced to just 20 (30.2 bit entropy). Because of this letters like *x, y, u, z* could be removed from being part of the password; thereby difficult words starting with them could be eliminated.
- In order to increase user involvement and memorability, a system could be created where after providing the users with a mnemonic they can be asked to draw a pictorial representation of it. This representation could be shows to them each time they login. This would not compromise security since the image does not mean anything to a stranger looking at it, but could be useful as a mnemonic aid to the users.

Below are some suggestions for designing a better experiment in future research;

- Studies involving participants from a wider range of demographics, so that the results can be generalized to a wider range of users.
- Involving more participants.
- Move to real-life settings outside the laboratory.
- Train participants on using mnemonic techniques like chunking and pronouncing.
- Screen participants based on short-term or long-term memory tests.
- Wait a longer time period between the creation and recall tasks to validate the results of the long-term recall of passwords across password policy conditions.

APPENDICES

Appendix A

Information Concerning Participation in a Research Study
Clemson University

Evaluating the Usability of Four Password Generation Schemes

Description of the Research and Your Participation

You are invited to participate in a research study conducted by Sanjaykumar Ranganayakulu under the direction of Dr. Joel Greenstein. The purpose of this research is to investigate the usability of four password generation schemes.

Your participation will involve being introduced to the research, signing this informed consent form, and using the password scheme assigned to you. After completing the first session of user testing, you will be asked to return after 5 to 7 days to complete a second set of password entry tasks and to provide feedback on the scheme. In both sessions you will also be asked to complete satisfaction and workload surveys.

The amount of time required for your participation will be approximately one hour for Session One and 30 minutes for Session Two.

Risks and Discomforts

There are no known risks associated with this research.

Potential Benefits

There are no known benefits to you that would result from your participation in this research. This research may help us to discover more usable and secure methods of generating passwords.

Protection of Confidentiality

We will do everything we can to protect your privacy. Collected data will be stored securely with access being limited to the investigators. Your identity will not be revealed in any publication that might result from this study.

In rare cases, a research study may be evaluated by an oversight agency, such as the Clemson University Institutional Review Board or the Federal Office for Human

Research Protections, which would require that we share the information we collect from you. If this happens, the information would only be used to determine if we conducted this study properly and adequately protected your rights as a participant.

Voluntary Participation

Your participation in this research study is voluntary. You may choose not to participate, and you may withdraw your consent to participate at any time. You will not be penalized in any way should you decide not to participate or to withdraw from this study.

Contact Information

If you have any questions or concerns about this study, or if any problems arise, please contact Dr. Joel Greenstein at Clemson University at 864-656-5649. If you have any questions or concerns about your rights as a research participant, please contact the Clemson University Office of Research Compliance (ORC) at 864-656-6460 or irb@clemson.edu. If you are outside of the Upstate South Carolina area, please use the ORC's toll-free number, 866-297-3071.

Consent

I have read this consent form and have been given the opportunity to ask questions. I give my consent to participate in this study.

Participant's signature: ___________________________________ Date: _________________

A copy of this consent form will be given to you.

Appendix B

PRE-TEST QUESTIONNAIRE

GENERAL

Participant: ______________________ (*This will be filled out by the test administrator*.)

Age: ______________________

Gender: □ Male □ Female

EDUCATION

1. Please select your academic level:

□ Undergraduate student
□ Graduate student
□ Other
(Please specify: ____)

2. List your major area of study: ____

COMPUTER EXPERIENCE

3. How long have you been using computers?

□ < 1 year □ 1-2 years □ 3-5 years □ > 5 years (Please specify) ________

4. How long have you used passwords?

□ < 1 year □ 1-2 years □ 3-5 years □ > 5 years (Please specify) ________

5. How many unique passwords do you have?

□ 1 □2 □ 3 □ More than 3 (Please specify the number) ________

Appendix C

Methodologies for remembering passwords*

*Source: Guide to Enterprise Password Management (Draft), NIST Special Publication 800-118 (Draft)

1. **Mnemonic Method:** A user selects a phrase and extracts a letter from each word (e.g., the first or second letter of each word), adding numbers or special characters or both.
   Example:

Phrase	Password
Please be my best valentine!	Pbmbval!
This is the worst car I have ever driven in my LIFE!	TitwcIhedimLIFE!
I am definitely your #1 fan.	Iady#1f.

2. **Altered Passphrases:** A user selects a phrase and alters it to form a derivation of that phrase.

   Example:

Passphrases	Alternate Passphrases
to be or not to be	2.be.0r.n0t@to0.bEE
Dressed to the nines	Dressed*2*the*9z

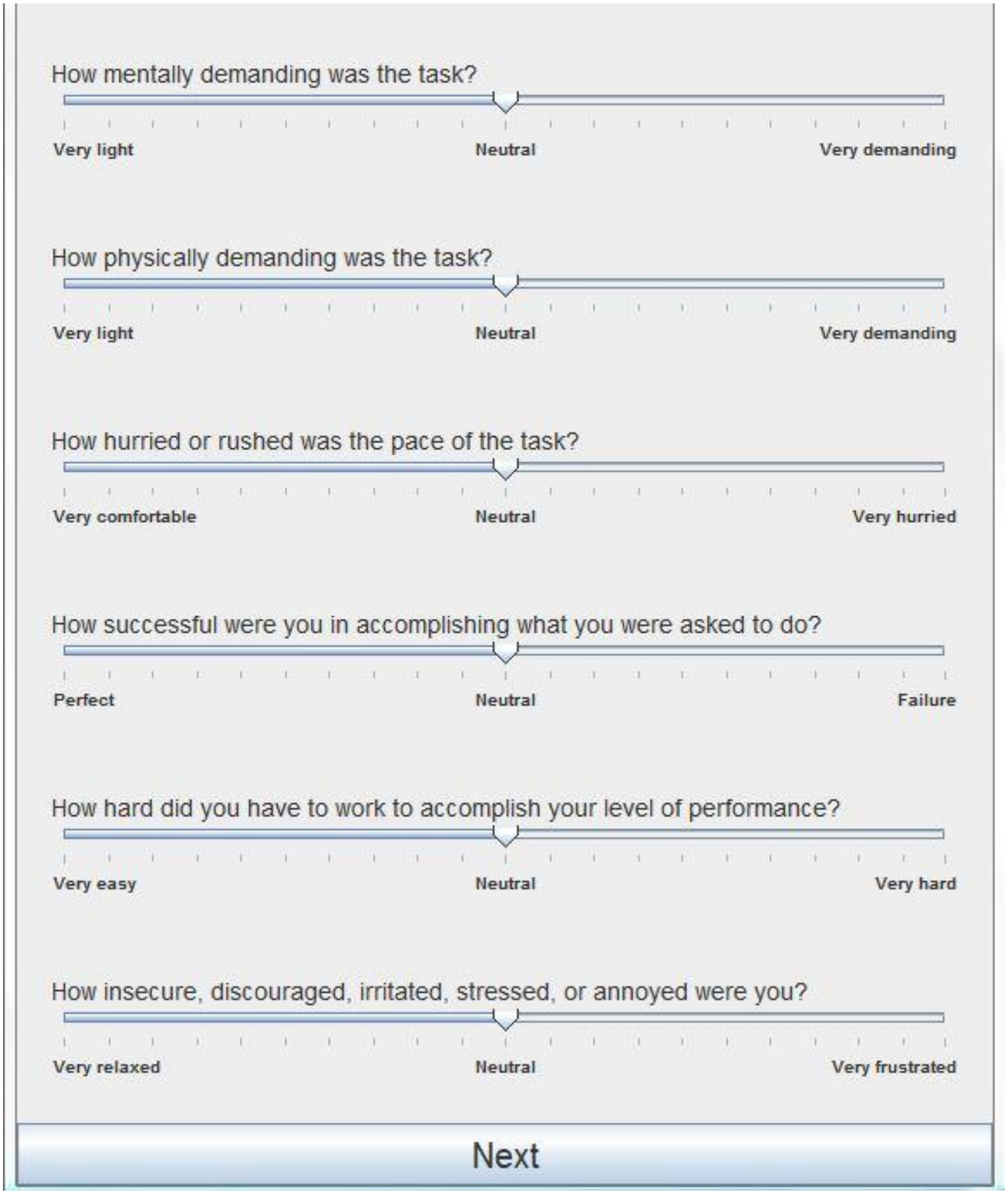
3. **Combining and Altering:** A user can combine two or three unrelated words and change
   various letters to numbers or special characters.

   Example:

Words	Password
“bank” and “camera”	B@nkC@mera
“mail” and “phone”	m4!lf0N3

Appendix D

NASA-TLX



How mentally demanding was the task?

Very light — Neutral — Very demanding

How physically demanding was the task?

Very light — Neutral — Very demanding

How hurried or rushed was the pace of the task?

Very comfortable — Neutral — Very hurried

How successful were you in accomplishing what you were asked to do?

Perfect — Neutral — Failure

How hard did you have to work to accomplish your level of performance?

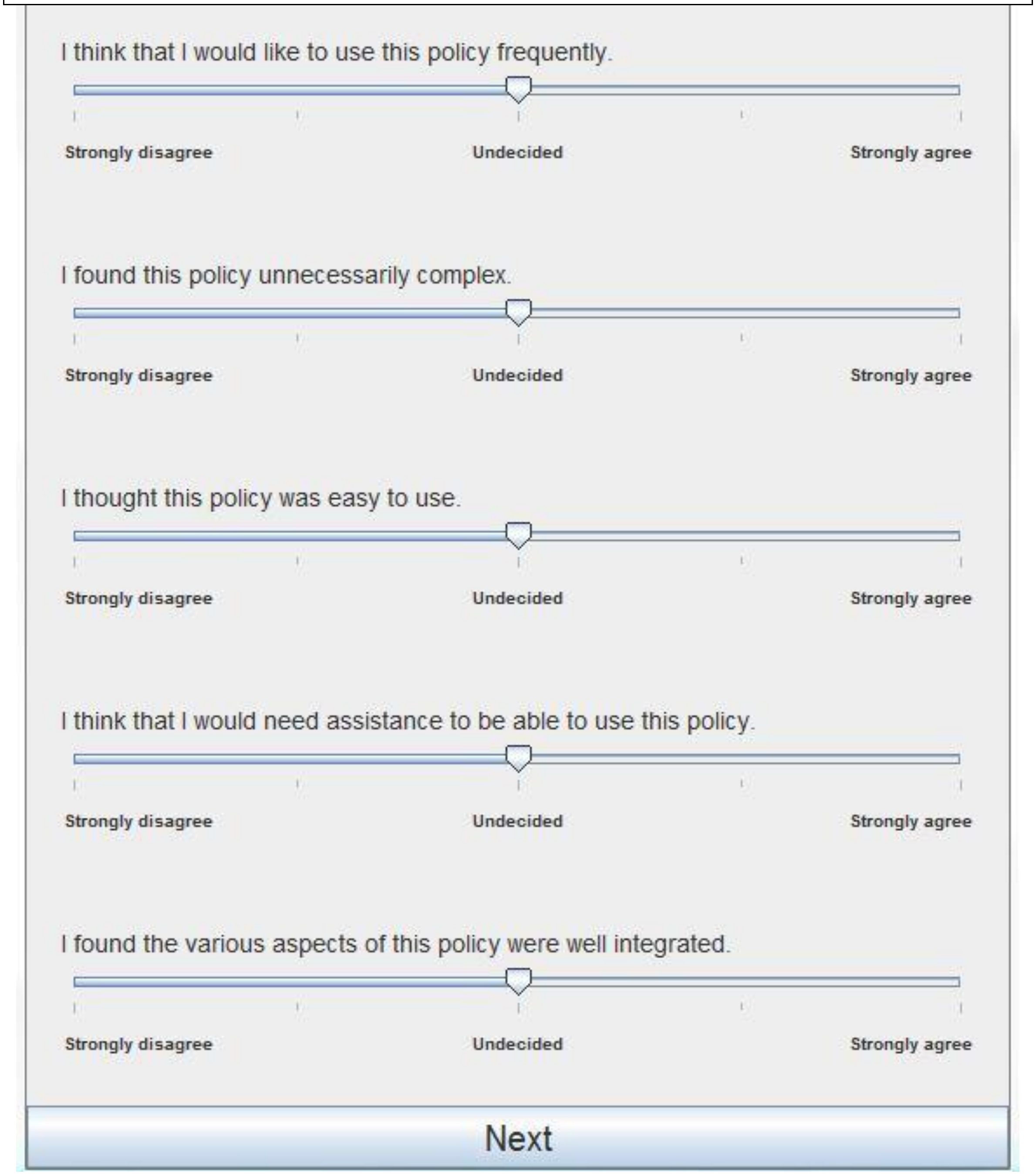
Very easy — Neutral — Very hard

How insecure, discouraged, irritated, stressed, or annoyed were you?

Very relaxed — Neutral — Very frustrated

Appendix E

System Usability Scale Questionnaire
System Usability Scale © Digital Equipment Corporation



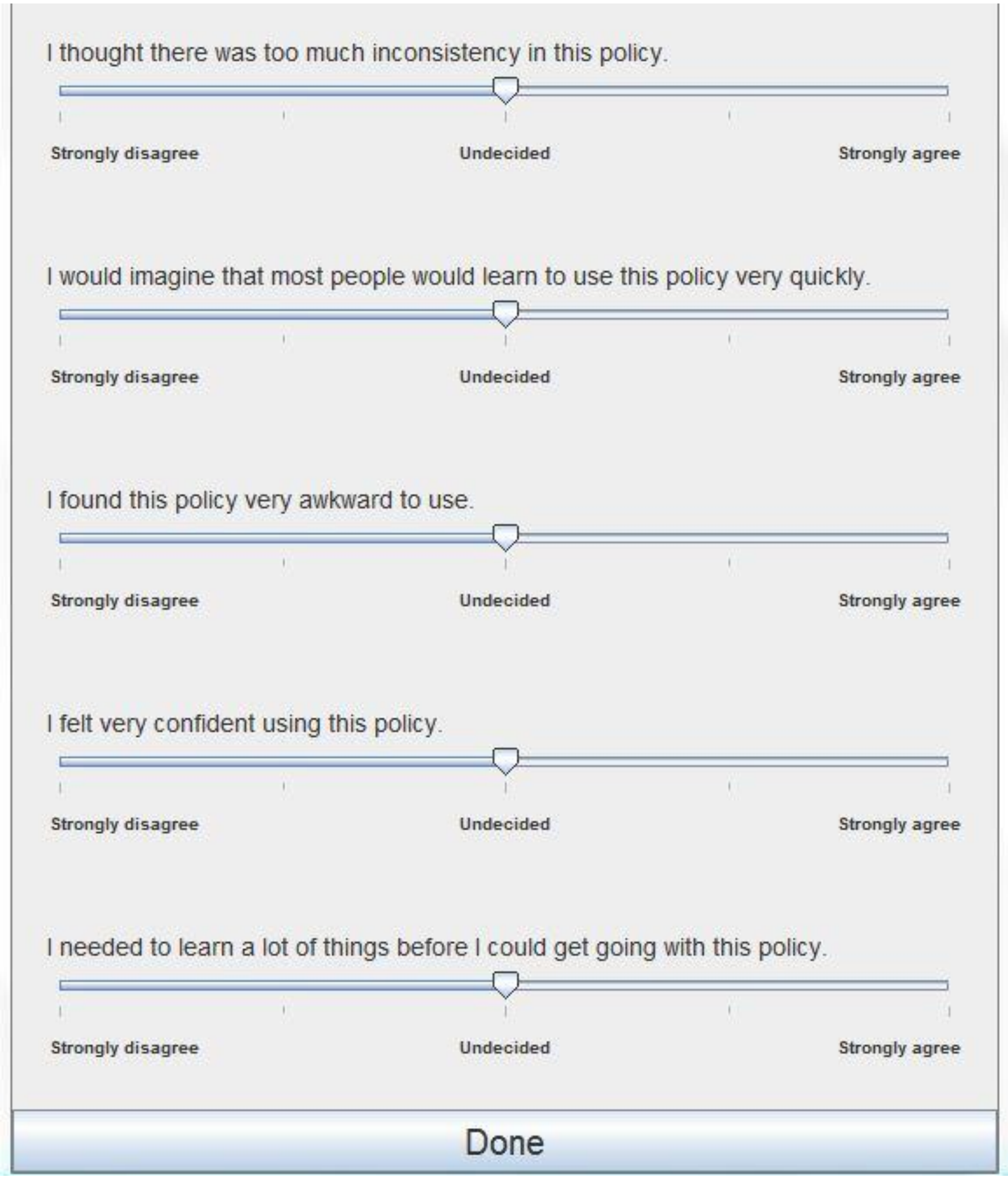
I think that I would like to use this policy frequently.

I found this policy unnecessarily complex.

I thought this policy was easy to use.

I think that I would need assistance to be able to use this policy.

I found the various aspects of this policy were well integrated.



I thought there was too much inconsistency in this policy.

Strongly disagree | Undecided | Strongly agree

I would imagine that most people would learn to use this policy very quickly.

Strongly disagree | Undecided | Strongly agree

I found this policy very awkward to use.

Strongly disagree | Undecided | Strongly agree

I felt very confident using this policy.

Strongly disagree | Undecided | Strongly agree

I needed to learn a lot of things before I could get going with this policy.

Strongly disagree | Undecided | Strongly agree

References

Brostoff, S., & Sasse, M. A. (2000). Are passfaces more usable than passwords? A field trial investigation. Paper presented at the Proceedings of HCI 2000, Retrieved from http://www.cs.ucl.ac.uk/staff/S.Brostoff/index_files/brostoff_sasse_hci2000.pdf

Bunnell, J., Podd, J., Henderson, R., Napier, R., & Kennedy-Moffat, J. (1997). Cognitive, associative and conventional passwords: Recall and guessing rates. *Computers & Security, 16*(7), 629-641. DOI: 10.1016/S0167-4048(97)00008-4

Burr, W. E., Dodson, D. F., & Polk, W. T. (April 2006). Electronic authentication guideline. NIST Special Publication, 800-63.

Damerau, F. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM, 7*(3), 171-176.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149-1160.

Jeyaraman, S.; Topkara, U. (2005). Have the cake and eat it too - infusing usability into text-password based authentication systems. Computer Security Applications Conference, 21st Annual, 1063-527, doi: 10.1109/CSAC.2005.28.Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1565274&isnumber=33214

Keith,M., Shao, B., & Steinbart, P. J. (2007). The usability of passphrases for authentication: An empirical field study. International Journal of Human Computer Studies, 65(1), 17-28. Retrieved from http://dx.doi.org/10.1016/j.ijhcs.2006.08.005

Klein, D. V. (1990). `Foiling the cracker': A survey of and improvements to, password security. Paper presented at the Proceedings of the United Kingdom Unix and open systems User Group (UKUUG) Conference, 147-54.

Kuo, C., Romanosky, S., & Cranor, L. F. (2006). Human selection of mnemonic phrase-based passwords. Paper presented at the Proceedings of the Second Symposium on Usable Privacy and Security, Pittsburgh, Pennsylvania. 67-78. Retrieved from: http://doi.acm.org/10.1145/1143120.1143129

Lehtinen, T. (2009, December 11). Angry birds. Retrieved from http://www.rovio.com/en/our-work/games/view/1/angry-birds

Leonhard, M. D., & Venkatakrishnan, V. N. (2007). A comparative study of three random password generators. Paper presented at the Electro/Information Technology, 2007 IEEE International Conference on, 227-232.

Morris .R & Thompson .K, (1978). “Password security: A case history,” Communications of the ACM, November 1979, 22(11):594 597.

Pond, R., Podd, J., Bunnell, J., & Henderson, R. (2000). Word association computer passwords: The effect of formulation techniques on recall and guessing rates. Computers & Security, 19(7), 645-56. Retrieved from http://dx.doi.org/10.1016/S0167-4048(00)07023-1

Proctor, R. W., Mei-ching Lien, Vu, K. -. L., Schultz, E. E., & Salvendy, G. (2002). Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods, Instruments, & Computers, 34*(2), 163-9.

Zviran, M., & Haga, W. J. (1993). A comparison of password techniques for multilevel authentication mechanisms. *Computer Journal, 36*(3), 227-37.