Narrative Passwords: Potential for Story-Based User Authentication

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with an

Major in Human Factors

in the

College of Graduate Studies

University of Idaho

by

Connor C. Hoover

Major Professor: Steffen Werner, Ph.D.

Committee Members: Rajal Cohen, Ph.D.; Gregory Turner-Rahman, Ph.D.

Department Administrator: Todd Thorsteinson, Ph.D.

December 2015

Authorization to Submit Thesis

This thesis of Connor Hoover, submitted for the degree of Master of Science with a Major in Human Factors and titled "Narrative Passwords: Potential for Story-Based User Authentication," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____ Date: _____

Committee Members: _____ Date: _____

_____ Date: _____

Department

Administrator: _____ Date: _____

Abstract

This paper explores alternatives to traditional alphanumeric passwords. Users are asked to remember random information in different cognitive modalities (i.e., words vs. images) to improve password retention and thus increase security. We present a summary of the literature on current approaches to passwords and the relevant literature on cognition and memory and propose a new "Narrative Passwords" authentication method. Randomly chosen verbal password elements are embedded in a short stories to make the information more memorable. Through several pilot studies we optimized the presentation of narrative passwords and which verbal elements are most memorable. In the main study, we compared Narrative Passwords to both a traditional, randomly generated alphanumeric password and a recently developed graphical password system, Composite Scene Analysis (Johnson & Werner, 2008). Our results indicate that Narrative Passwords are not as memorable as similarly graphical passwords but the systems could be combined to increase their effectiveness.

Table of Contents

List of Tables

List of Figures

Introduction

1.1 Computer Security

Passwords are simple, effective, and low cost security measures that have been used for centuries. The basic premise of a password is that access to stored information requires knowledge of a secret. This secret could be a word, a phrase, or a string of numbers, but it is not limited to the domain of language and writing. Currently, passwords are the most common type of electronic authentication system as well as the most common "cognitive" authentication system: the ability to authenticate relies primarily on the user's cognitive abilities. In an ideal world, a password or code phrase takes only seconds to create and all the user has to do is remember it. This seeming simplicity comes with two major drawbacks. First, if a password is compromised, the system provides no additional security measures. The second drawback is that the system is reliant on the user being able to remember the password; if they cannot do so then they cannot access the information (Adams, Sasse, & Lunt, 1997; O'Gorman, 2003).

Advances in the field of cybersecurity have been largely driven by computer science and computer engineering considerations. The user of these systems is treated more or less as an afterthought. By passively resisting the adoption of principles of human behavior, cybersecurity has failed to acknowledge a fundamental aspect of security systems: the user is frequently the weakest link. The most common method of hacking a password-secured system by exploiting this weakness is called "phishing" The basic technique involves using fake websites, emails, instant messages, etc. to trick the user into revealing their password information. No attempt is made to break the software or hardware; it is much easier to

break the wetware (i.e. the user) (Josang et al, 2007). Sometimes, it is not necessary to trick the user into revealing the password at all. Many users reuse passwords or choose predictable passwords based on personal or easy-to-remember information. The "strength" of a password (*H*) is determined by a formula: $H = \sum_{i=1}^{L} log_2 n_i$ where *L* is the length of the password and n_i is the number of potential characters that could be chosen for that position. For a password where each character comes from the same pool of options and is equally likely to be chosen (a random password) the formula can be simplified to $H = L(log_2 N)$, where *L* is the length of the password and *N* is the number of possible password characters. This formula produces *H,* the number of bits of entropy, a measure of information randomness based on the number of possible character combinations. A higher bit number indicates more potential random combinations and therefore a more secure password. However, when users choose non-random passwords they reduce the actual strength of the password compared to the theoretical strength. Users do this because random information is difficult to remember. Short, meaningful, and frequently used character combinations are simple for users to remember, but provide minimal security. What is optimal for the user and optimal for the system are entirely at odds (Burr, Dodson, & Polk, 2006; Florencio & Herley, 2007).

1.2 Cognitive Passwords

The tradeoff between memory and security is called the "password problem" and shows the necessity of considering human cognitive capabilities when developing password system protocols. The development focus needs to be placed on strengthening the weakest part of the system. Cognitive sciences provide a wealth of data that can inform the

development of these user-centered systems. There have already been attempts at creating user-centered password systems. Visual-spatial passwords are one increasingly popular development based on these principles (Biddle, Chiasson, & van Oorschot, 2011; De Angeli et al, 2005; Renaud & De Angeli, 2004). For example, many modern Smartphones offer a Pattern Lock password system (Andriotis, et al. 2013). The user connects dots on a grid (3x3 to 9x9) by drawing lines between them and creating a pattern. Other commercialized graphical password systems include Passfaces™ (Passfaces Corporation, 2005). As the name implies, Passfaces uses images of faces as the authentication mechanism. The user has to recognize and identify the correct series of faces in order to access the system (Brostoff & Sasse, 2000). Another visual password prototype that uses a similar graphical recognition authentication process is Composite Scene Authentication (CSA), which can help users encode a greater number of bits of entropy (46.5 bits for CSA compared to roughly 19 bit for the grid) (Andriotis, et al. 2013; Johnson & Werner, 2006, 2007, 2008). Possibly the most important feature of these password systems is they rely on recognition instead of serial recall. For CSA, users are initially presented with a picture, composed of randomly selected smaller images elements, which will acts as the password. Later on they were asked to select the correct images out of a larger set that contains both the correct images as well as distractors, which were not presented as part of their password picture. The length of a regular password is equivalent to the number of images or image elements to be recognized through the presentation of successive authentication screens. Each target is individually displayed on the authentication screen among a number of distractors, so the information entropy for each element is determined by the number of distractors. Relying on recognition

of the correct image allows users to successfully identify the correct image more accurately than in systems that requires them to reproduce an image from memory (Shepard, 1967). Previous research has indicated the existence of a "picture superiority effect", meaning that pictures are more easily remembered than words (Goldstein & Chance, 1970; Shepard, 1967; Standing, 1973). This finding has helped drive the creation of visual passwords. However, using visual passwords does have drawbacks. Blind or low vision users would be unable to use these systems, which is contradictory to principles of universal accessibility. Furthermore, because the composite scenes are generated from randomly selected images the composite image lacks cohesiveness. Preliminary data indicates that if users perceive the relationship between parts of an image as meaningful (not random), then they are more successful at remembering those images individually (De Angeli et al, 2005; Renaud & De Angeli, 2004; Johnson & Werner, 2008).

1.3 Verbal Material and Passwords

One example of a security authentication system that requires the use of verbal material is a system where users provide answers to "Security Questions" ("What was your favorite pet's name?", "What was your mother's maiden name?") in case they forget their password and need to reset it. The weakness of these questions is that they usually refer to something about the user's history, which investigation on the part of a hacker could discover (Keith, Shao, & Steinbart, 2007; Kurzban, 1985). The flaws in this system were made apparent to the public in October of 2014 when hackers broke into the iCloud backup accounts of a large number of high profile women, stole photographs stored there, and

published them online. This egregious breach of privacy was possible, according to some security experts, because Apple still used Security Questions as part of their password system. Because these women were in the public eye, it was particularly easy for hackers to research them and guess their responses to the Security Questions backup system, though in an era of social networking it is not only public figures who are at risk for this kind of data mining. This allowed them to reset the user's password and access the system (Cubrilovic, 2014). Using information that can be learned by malefactors as a security backup is the equivalent of choosing a low entropy, easy-to-guess password to protect the system. The fault lies in the fact that the system implicitly encourages the user to make these choices for ease of use, rather than choosing safer, more difficult passwords and answers.

Another option for password creation that uses verbal material is passphrases. A passphrase is a longer string of characters that acts as the password. A passphrase can consist of multiple randomly chosen words ("correcthorsebatterystaple") or some kind of meaningful sentence ("therainінspainfallsmainlyintheplains") (Keith, Shao, & Steinbart, 2007). Increasing the number of characters increases the password entropy, while at the same time phrases are easier to remember than random numbers and letters. However, there are a few drawbacks associated with the use of passphrases. Research has shown that while the phrases are as memorable as a user generated or randomly generated password, their increased length creates a higher risk of orthographical errors during creation and authentication, causing users to perceive them as more difficult to use (Keith, Shao, & Steinbart, 2007). Furthermore, user generated passphrases suffer from the same weaknesses as passwords; they can be predictable, making them vulnerable to "dictionary

attacks" where hackers attempt to authenticate with a large number of common words and combinations (Kuo, Romanosky, & Cranor, 2006).

Frequently, security systems will require users to generate a password within a certain set of restrictions in an effort to compromise between memorability and security. This forces the users to create new strategies for password generation such as replacing letters with symbols, using the first letters of words in a sentence, or entering characters based on a pattern of keys on the keyboard (Brown et al., 2004; Vu et al., 2007; Zviran & Haga, 1993). While these techniques do improve memory for user generated passwords, they do not necessarily improve security. Since the passwords are still chosen by the users, those choices are still relatively predictable. When replacing letters with numbers for example, users tend to make common replacements, such as replacing "E" with "3" (Brown et al., 2004). Keyboard patterns and first letter selection are more effective techniques than letter replacement, but still have certain drawbacks. Using the first letters of a sentence can be subject to the same problems as passphrases: generating a long enough password can lead to entry errors. It also requires the user to remember a much longer phrase which introduces the possibility of forgetting or omitting one or more words (Vu et al., 2007). Keyboard patterns are more vulnerable to shoulder-surfing, where a hacker spies on the user while they type their password, as well as brute force algorithms that include common patterns of keypresses. Finally, imposing restrictions on user-generated passwords does not necessarily result in more secure passwords (Zviran & Haga, 1993). If all the passwords have the same requirements, then the theoretical entropy space is reduced. Users may also be unable to fully use one of the previously mentioned strategies which could result in more

difficultly remembering the password, or they may attempt to circumvent the restrictions and use low entropy passwords that only barely meet the requirements and are highly predictable.

1.4 Biometrics and Token-Based Systems

Apart from cognitive authentication systems there are several other types of security systems. Biometric systems such as fingerprint scanners and retinal scans are becoming more common as the technology required for them becomes easily available. Many current generation smartphones have optional fingerprint systems. The primary drawbacks to these systems are fidelity and accessibility for disabled user groups. While the resolution on biometric systems is greater than it once was, it is still possible to fool the system with a similar enough imposter. The issue becomes a tradeoff between accuracy and usability: the more accurate the system becomes the safer it is but the more discriminating it becomes as well. If the user suffers damage (say to their fingertip) they may become locked out of their system either temporarily or permanently (Prabhakar, Pankanti, & Jain, 2003). Furthermore, not all users have sufficient use of their hands or eyes to be able to easily access a biometric system. Typically, once compromised biometric systems are extremely difficult to reset. Users can only reset a fingerprint password ten times at most or a retinal system once before they are out of alternatives (Jain, Bulle, & Pankanti, 1999).

Another option for authentication is token-based authentication. A token is a physical object used to access a secure system. Most commonly these tokens are keycards, such as the room keys used by many hotels, although other kinds do exist, such as USB sticks

that contain long encryption keys. Many token based systems use multi-factor authentication, i.e. a token and a password, like an ATM card and PIN (Brainard et al., 2006). The problem of course is that since tokens are physical objects, they are vulnerable to theft and loss. The advantage of the multi-factor system is that if just the token is lost, or just the password compromised then the system is still protected. However, securely replacing a lost token may be difficult or time consuming for the user, and they remain locked out of the system while they do so.

There will always be a need for cognitive authentication systems that are user centric, accessible, and inexpensive. Cognitive passwords allow users to choose the modality and input style that suits them best, as well as provide additional security and redundancy for one of the other systems mentioned above. If a user wished to access a system remotely, where a biometric scanner or token reader was not available, a redundant cognitive password could facilitate that. A cognitive password that uses verbal material can be entered with voice recognition software for users who have difficulty with a standard keyboard interface. Visually impaired users can still access systems that do not rely on visual material, such as security questions that can be read aloud. Non-reading, low literacy, or hearing impaired users can still make full use of graphically represented material to access a system. Providing a wealth of accessibility options improves the experience for a variety of different user groups with minimal sacrifices to the overall security.

1.5 Creating a Memorable and Usable Cognitive Password

The primary criterion of a usable cognitive password is the ability of the user to remember the information they need to authenticate into the system regardless of the modality of the information itself. Understanding how users remember information and which memory processes are most effective will necessarily inform the development of any cognitive password system. The most important distinction in types of memory is between recall and recognition. Recall memory is the ability to recall events or information from episodic or semantic memory. These events could range from the learning of a new word, after which usage of that word would constitute an instance of recall, to recollection of an entire sequence of events from childhood. Recall is commonly used in psychological tests in a variety of ways. Within recall there are distinctions relating to how it is used in testing. Free recall is the ability to retrieve memories without any cue present. An example of this would be an experimenter asking a subject to recall an event that happened previously. Cued recall involves the use of an associate that helps trigger recall of the desired item. An example of this type of recall is the learning of paired associate words. In this sort of task, a participant would learn a pair of words, such as “table/chair”, and then later they are presented with “chair” and must recall its associate “table”. In these instances, the participant is using a cue to aid in the recall of the desired stimulus. The final type of recall task is serial recall. In a serial recall task, a participant is required to remember items in the order they learned them. For example, the participant may learn a list of words or numbers and the recite them in the same order as they were on the list (Yonelinas, 2001, 2002).

Recognition is functionally defined as the ability to distinguish previously encountered stimuli from new stimuli (Yonelinas, 2002). As opposed to recall, in a recognition task the stimulus to be retrieved from long-term memory is present. The participant simply needs to make a judgment about whether or not they have encountered that stimulus within the context of the experiment before. Most participants show greater accuracy on recognition tasks versus recall tasks. There are multiple theories about the process of recognition and why it shows greater accuracy. One of the most popular theories is Generation-Recognition theory. Generation-Recognition theory states that the process of retrieval involves, after the initiation of the retrieval, the generation of items from long-term memory that are semantically related to the desired stimulus, and then the recognition of the appropriate item from among those generated. This two-stage process accounts for the gap in accuracy from recall to recognition; recall uses both parts of the process, while recognition skips the first stage altogether (Medina, 2008; Ranganath et al., 2004). In addition to being more accurate, recognition is a faster process than cued recall, which has been demonstrated experimentally (Nobel & Shiffrin, 2001).

Another theory about what makes recall and recognition different is the theory of encoding specificity. The encoding specificity principle states that retrieval depends on cues encoded with a stimulus being recognized concurrently. A study by Tulving and Thomson (1973) found instances in which recall proved more accurate than recognition, as long as there was a cue present that had been encoded with the information during recall. Encoding specificity accounts for other memory phenomena such as state dependent memory, where information is better recalled in the context in which it was learned.

Because of the improvement in memory associated with cued recognition memory, a highly usable cognitive password system should allow the user to recognize the correct way to authenticate rather than requiring them reproduce it from memory alone. This principle is already used in some cognitive authentication systems such as the graphical authentication system discussed earlier.

Of concern for recognition based systems like CSA is the "list length effect" that has been observed in recognition memory tasks. This effect appears as greater accuracy in recognition when the stimuli are presented in short lists at study compared to long lists. The concern relates to potential interference between items during encoding which produces confusion later during testing, swapping of two semantically or visually similar stimuli, or duplication, or deletion. This effect could limit the number of items that could be effectively encoded by the participant and used for authentication, diminishing the overall entropy of the system. However, recent studies have found that when variables such as boredom, retention and distraction intervals, and rehearsal are controlled, the list length effect only occurs for certain stimuli (Kinnell & Dennis, 2012). Both word pairs and pictures appear to be discriminable enough that they do not interefere with one another while encoding. This finding provides support for the idea that pictures or text could be used effectively in recognition based authentication. While commercial graphical authentication systems exist and research on using graphical systems has been published, little work has been done on exploring similar text based systems and the situations in which an alternative modality may be useful. To further explore this idea, it is necessary to develop a better understanding of how humans comprehend and remember text.

1.6 Text Comprehension

Modern theories of text comprehension state that readers create a mental model of the text they are reading by breaking it down into multiple layers that are encoded hierarchically and interact to produce measurable memory effects (van Dijk & Kintsch 1983; Kintsch 1998). The first layer is the surface structure of the text. This includes graphic representations of words, syntax, paragraph structure and other physical features. Testing memory for surface structure would involve participants recreating the visual properties of the text; paragraph structure, length, word count, etc.

The second layer is the propositional content of the text. A proposition is the primary unit of information in language. According to Kintsch (1974), text can be decomposed into sets of nested organized propositions, which are defined as multi-element sets of lexical items. 'Mary bakes a cake.' Can be organized as the following proposition (BAKE, MARY, CAKE) where BAKE is the predicate and MARY & CAKE are arguments (subject and object). In this model, the uppercase words do not just refer to the surface structure equivalent that the subject reads, but to the entire semantic memory entry for that word in the reader's memory. Propositions can be nested and organized into categories and sub-categories that reflect the structure of the original text; thus it is possible to attempt to derive shallower encoding layers from deeper ones, but the transfer is not perfect. Testing memory at this level of encoding would involve recall or recognition of propositions that did or did not occur in the text (was there a cake? Who baked the cake?).

The final layer in the mental representation is the reader's perception of the meaning of the text (Radvansky & Zachs, 1991; Radvansky, Spieler & Zacks 1993). The propositional content is compared to existing semantic and episodic memory entries for the lexical content and incorporated into a "situation model" of the content of the material. To test a reader's memory at this level, the participant may be asked to describe the situation presented in the text as accurately as they remember it. Research shows that of the three levels, the situation model is remembered best over long delays. This was demonstrated by Bartlett (1932) with his "The War of the Ghosts" study, in which the first two layers of the model are lost when a participant is asked to recreate text material over long retention intervals (days, weeks, and years after the initial reading). The situation model however, was much more robust than the syntactical style or propositional content; certain propositions were rearranged or dropped entirely while the basic situation remained. However, any errors that occurred when the participant initially translated the propositional content into situational content, changing the main characters from "two men" to "two brothers", for example, were equally robust.

Other studies have supported the idea that the situation model has the largest influence on memory for text (Fletcher & Chrysler, 1990; Radvansky & Zachs, 1991; Radvansky, Spieler & Zacks 1993). However, all three levels can be used to support memory and are more effective when combined than alone. Fletcher and Chrysler (1990) demonstrated this by having participants read sentences, then testing recognition for the correct sentence compared to distractors that were inconsistent on one of the three levels. Another study by Radvansky and Zacks also investigated recognition of previously read

sentences. In this instance, the encoded sentences were manipulated as to whether they described one object in multiple locations or multiple objects in the same location. They concluded that recognizing the multiple object sentences was faster than recognizing multiple locations. They attributed this effect to the idea that it is easier for a reader to develop a situation model in which multiple objects inhabit the same location than it is to develop a model wherein the same object inhabits multiple locations (Radvansky & Zacks 1991; Radvansky, Spieler & Zacks 1993).

1.7 Narrative Passwords

Using the information garnered from text comprehension, recognition memory, and verbal material it is possible to create a new cognitive authentication system called a "narrative" password system. A narrative password uses a randomly generated short story as the system password. The system is conceptually similar to the CSA password system mentioned before, although it differs in a number of key ways. Using verbal material provides a number of benefits which a system based on visual material could not. Verbal material structured as a story provides meaningful context for the information while still allowing for the random generation of the story itself, preserving the meaningful word choices while eliminating the weakness of user generation. Different elements of a story, such as the names of different characters, the locations in the story, etc. can be randomly selected to produce a combinatorial password with high potential entropy, in which each element is individually meaningful. Previous research indicates that different types and attributes of words play a role in how memorable they are: nouns are generally more

memorable than adjectives and adjectives are generally more memorable than verbs (Gentner, 1981; Pavio, 1986; Shepard, 1973). When nouns and adjectives or nouns and verbs are paired together memory experiments show that these types of cued recall scenarios improve subjects' memory (Gorman, 1961; Kersten & Earles, 2004; Pavio, 1963). Pairing words to be remembered would allow users of narrative password systems to benefit from context, recognition, priming, and cued recall. All of these factors have well supported cognitive effects on memory performance. In addition, individual words can be assessed as to how specific, novel, or abstract they are, for which rating scales already exist. Words that are novel or distinctive are easier for participants to recognize when placed alongside other novel distractors during memory tests than recognizing non-novel items among other non-novel distractors (Kishiyama & Yonelinas, 2003). Similarly, words with a low frequency of use in common language attract greater attention than commonly used words, which produces a larger memory effect (Malmberg & Nelson, 2003). Each of these dimensions can help determine how easily an individual word may be encoded (Gorman, 1961; Pavio, 1963, 1967). Therefore, it is possible to optimize the selection of words that will serve as elements of a narrative password.

The structure of a story potentially offers some unique memorability advantages over visual material. Stories naturally include the repetition of names or other specific words. Thus, rehearsal, known to be extremely beneficial for memorization, is inherent in the medium. In addition, stories produce specific mental images through description. For example, the protagonist in the story can be described in a paragraph, which creates a mental image; after that, it only takes the name of the protagonist to cue recall of that

mental image. Mental imagery has also been shown to be beneficial for encoding information. Using multiple strategies to encode information is more effective than using only one. Therefore, the possibility of combining rehearsal and imagery in one story give the narrative password system a strong advantage in memorability over many other systems (Gorman, 1961; Pavio, 1986; Seamon, 1972). The advantage for multiple modalities also means that stories and images could be incorporated together into a single system. However, before examining the potential for a multiple modality system it is prudent to examine and test both systems in isolation to develop a greater understanding of their relative strengths and weakness so that a combined systems could be deliberately and conscientiously designed.

The nature of a story also allows for words and phrases to be emotionally valenced. In isolation, most words do not have a particularly strong emotional valence. However, when used as properties of a larger story, words that normally have no emotional connotations can be imbued with such. Negative valence words have been shown to be more memorable than neutral valence words, suggesting that sad or tragic stories may make better passwords (pending further investigation) (Maratos, Allan, & Rugg, 2000).

Verbal material can be presented in visual or auditory formats, making narrative passwords accessible to vision or hearing impaired users and low literacy users. Authentication in a narrative password system could also take advantage of recognition memory by presenting the user with a similar series of authentication screens, where the correct element that appeared in the story is selected from list of distracters. However, it

would be possible to present a greater number of words per-screen compared to pictures per-screen, increasing the password entropy. This would also reduce the vulnerability of the narrative password system to the practice of “shoulder surfing”. Words are still easily discriminable at small text sizes while pictures become more difficult to recognize quickly. Finally, words can be organized alphabetically which is standardized and can speed up visual search, while there is not a similar standardized way to organize pictures.

A question that must be answered when designing a narrative password system is: what importance should be given to each of the individual properties of different words and text? While all of the previously mentioned studies have found various effects in isolation, no study has examined how those effects interact with each other or the relative contributions to memory they might have. It is unclear, for example, whether part of speech, frequency of use in language, or emotional valence would have the largest effect (or any discernable effect at all), in the context of a narrative password. While the existence of these effects supports the hypothesis that narrative passwords could make for an effective cognitive authentication system, trying to examine all of these effects during the initial development of the system would create far too many variables for a practical study. They represent optimization issues that can be addressed in the future to fine tune the effectiveness of the system. During the pilot studies conducted to investigate the potential of narrative passwords, development and selection of password elements was driven largely by practical concerns and needs. The results of each pilot study informed the next studies, creating a gradually evolving system. Therefore, not all of the theoretical perspectives mentioned here were taken into account during development, but it was not necessary to

do so as the pilot studies were interested in the gross effects and development potential. It would not make sense to use fine detail adjustments before making the coarse adjustments necessary in the first steps of system development.

Pilot Studies

2.1 Initial Pilot Study: Selecting Password Elements

To study the feasibility of a narrative password system a series of pilot studies were conducted each of which examined different aspects of the narrative password system. The first pilot study was conducted with 19 participants. The purpose of this initial study was to determine what elements of a short story made the most sense as password elements. To this end, a short story selected from an online data based was used and distributed to the participants. Below is an excerpt from the story (Fisk, 2011).

Ring Worlds

> Sir Charles Wilton had just poured himself a glass of brandy and flipped open a book he'd been looking forward to reading, when a sudden whooshing sound made him look up in time to witness a demon materializing in the library… – Peter Fisk

After being allowed to read at their normal pace, each participant was given a piece of paper and asked to summarize the story they had read. After these summaries were finished, a second sheet of numbered lines was distributed. On the lines participants were asked to list in order what they felt were the most memorable aspects of the story they had summarized.

The frequency of items listed in both the summary and ordered lists was tabulated and the 16 most frequently mentioned items were identified as potential password elements for this particular story. The items most frequently remembered, in order, were:

1. The gender of the protagonist	2. The identity of the antagonist
3. An object given to the protagonist	4. An object given to the antagonist
5. The protagonist's drink	6. A feature of the object given
7. Another feature of the object	8. The color of the object
9. The color of a different object	10. The protagonist's first name
11. The protagonist's last name	12. The location act
13. Location of the second act	14. A smell
15. The name of a minor character	16. The final line of the story

Table 2.1. List of the 16 most commonly identified items within the story

2.2 Study Two: Creating New Stories and Story Presentation

In a second follow up study we expanded on these results which consisted of testing 42 additional participants in a single, 45 minute experimental session. In an attempt to examine how effective narrative passwords are under a variety of conditions, the study examined three different variables. The first variable was story version. To see if randomizing the story affected participants' memory, two alternative stories were created by substituting one randomly selected alternative element for each story element. Participants in four of the experimental conditions received a new story, and participants in the other four received the original. In the first part of each session, the

story was distributed to each participant, after which they began reading at the same time the stopwatch was started. The average reading time for the story, both versions of which were 981 words long for the original and 1000 words long for the random version was 4:40 minutes (4:21 minutes for the original version, and 5:00 minutes for the random version).

The second variable investigated was rehearsal. To examine the effects of distraction and rehearsal on participants' memory, two sets of tasks were created. Participants in four of the conditions received a summary plus distractor task, and the rest received only a distractor task. In the summary conditions, immediately after all participants finished reading their assigned story they were asked to spend ten minutes writing a summary of the story. Participants were encouraged to be as detailed as possible. After the summary, participants were asked to spend five minutes writing a list of elements from the story that they thought they could most easily remember. Finally, participants were asked to spend five more minutes working on a distractor task, in this case a Sudoku puzzle. In the non-summary conditions, participants spent 20 minutes working on Sudoku puzzles after finishing the story. After finishing either of these tasks, participants the participants were tested with a recognition test. This variable was meant to evaluate how much impact the ability to rehearse the story information had on memory during the memory testing phase.

The third variable investigated as part of this study was order of question presentation. During the memory testing phase, participants were provided a

recognition test to assess the amount of information they had retained. The test contained 16 questions, one for each manipulated story element. Each question was a multiple choice question where participants selected the correct answer about the story they had read out of all of the potential answers that had been created. As a between groups variable, half of these tests presented the questions in an order that followed the progression of the story (questions about elements that appeared early in the story were asked first, and elements that appeared at the end were asked at the end). The other half of the tests asked the questions in a random order. This variable was meant to assess whether or not the authentication screens of a narrative password system could be used to provide further recognition cues to the user. Participants were randomly assigned to one of the eight conditions, with a minimum of four participants per condition, resulting in a 2 x 2 x 2 mixed design.

The results of the study are summarized in the tables below. For the participants in each group the percentage of correct answers was recorded. Correct answers were then used to calculate the bits of entropy retained during the study. To do this, the number of correct answers on the recognition test was entered into the formula for entropy as the length with the number of distractors as the number of character options for each participant. This produced the bits of entropy retained, which was averaged for each group.

Condition Means	Average Bits of Entropy Remembered			
Summary	44.69			
Non-Summary	39.76			
Ordered Test	44.97			
Random Test	39.48			
Original Story	46.86			
New Story	37.59			
Group Means	Summary		Non-Summary	
Average Bits of Entropy Remembered	Ordered Test	Random Test	Ordered Test	Random Test
Original Story	52.79	44.79	47.59	42.27
New Story	46.59	34.59	32.91	36.27

Table 2.2. Average bits of entropy remembered for each group.

Group Means	Summary		Non-Summary	
Average % Questions Correct	Ordered Test	Random Test	Ordered Test	Random Test
Original Story	80%	73.75%	78.13%	69.79%
Random Story	76.56%	57.78%	55.21%	60.42%
Condition Means	Average % Questions Correct			
Summary	72.02%			
Non-Summary	65.89%			
Ordered Test	72.47%			
Random Test	65.44%			
Original Story	75.42%			
New Story	62.49%			

Table 2.3. Average percentage of questions correctly answered for each group.

An ANOVA of story version, test order, and summary showed a main effect of story type on the results with participants showing greater recognition accuracy for the

original version: A 13% performance difference or ~ 9 bits of information difference between the groups ($F(1,41) = 10.00$, $p = 0.003$, $MS = 0.20$). However this effect was not found in later studies after the random version was edited into a more generic skeleton for greater readability. Two trends of question order ($F(1, 41) = 2.00$, $p = 0.16$, $MS = 0.04$) and summary condition ($F(1,41) = 2.50$, $p = 0.12$, $MS = 0.05$) were found but were not large enough to be significant: A 7% difference, or ~5 bits of information. However, this may have been due to the sample size and the trend was still taken into account for subsequent studies. Based on the results found here, all later studies asked recognition questions in the ordered they appeared in the story as that did appear to influence performance. While having participants write a summary did predictably improve performance, it would not be practical to have users of a narrative password write summaries during the encoding procedure. For the purpose of greater face validity, subsequent studies did not include summaries.

2.3 Study Three: A New Way of Studying Recognition Memory

The third study added a novel procedure for examining memory and information in short stories. This procedure was called step-down recognition. Of the original 16 elements in the random story, 11 were used in the second study. These 11 were chosen because they had the largest numbers of potential alternatives (e.g. the gender of the protagonist was not used because of the low number of alternatives). Next, a set of lists of the alternative words for each of the 11 elements were created. The largest of these lists contained 1024 elements, with each subsequent list being reduced in size by half

until only one element remained (the element which was used in the story). This procedure replaced the summary and recognition test of the first experiment. The story was also modified to be slightly shorter: 861 words, for a shorter average reading time of 4:05 minutes. After reading the random story, participants experienced a five minute distraction interval (Sudoku puzzles were replaced by 4th grade math problems because a number of the participants were confused by the rules of Sudoku and we wanted a task everyone would be familiar with). Following the distraction interval the participants were asked a series of 11 questions, each one relating to one of the randomized elements within the story. After the experimenter asked a question, the participant first attempted to give a free recall response as an answer. Next, the participant was instructed to view the first list of potential answers to that question (starting with the 1024 item list). The participant was forced to choose one answer from the list, and they were asked to give an estimate of their confidence in that answer (0% - 100%). If the participant was not 100% confident in their answer, the experimenter would show the participant the next list for that question, containing half the number of distractors (1024 items became 512 items). The participant repeated choosing an answer and providing a confidence estimate. This continued until the participant was 100% confident or until they reached the list with only one element. To incentivize the participants not to answer incorrectly a point system was implemented where points were awarded for correct answers and deducted for incorrect answers. Correct and incorrect answers were recorded, as well as the list size of the correct answers. Bits of information remembered were calculated using the entropy equation again: List size of

the 100% confident answer was used as the number of possible elements and the bits of entropy for each correct response were summed for each participant.

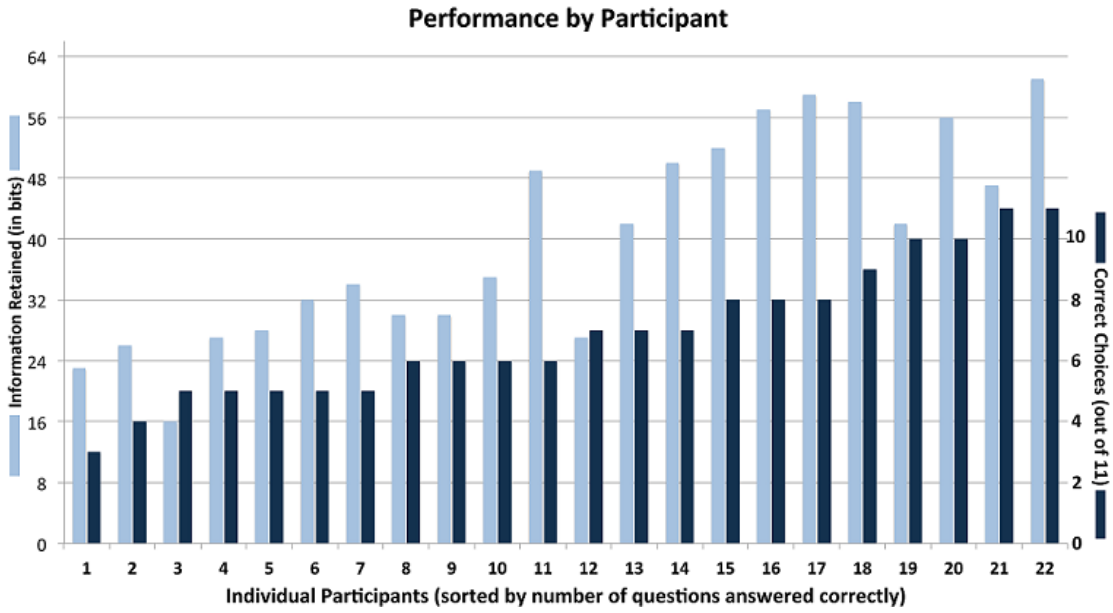


Figure 2.1. Performance of each participant in the second study. The light blue bars indicate the total bits of entropy retained by each participant, and dark blue bars represent total correct answers.

The results of the study indicated that the step-down recognition procedure was a more sensitive measure of performance than binary correct/incorrect scoring. While the two measures have a correlation coefficient of 0.78, the average bits of entropy for each participant took into account at what list size each participant correctly identified the target item. This measure then not only indicated how frequently the participant answered correctly, but also their degree of confidence and ability to discriminate between distractors. For example, as shown in Figure 2.1, participants 21 and 22 both answered all 11 questions correctly. Participant 22, however, scored nearly 20 bits higher on retained information indicating they were much more confident and

answered at higher list levels. The second study was also consistent with the first in that participants remembered an average of 40 bits of information.

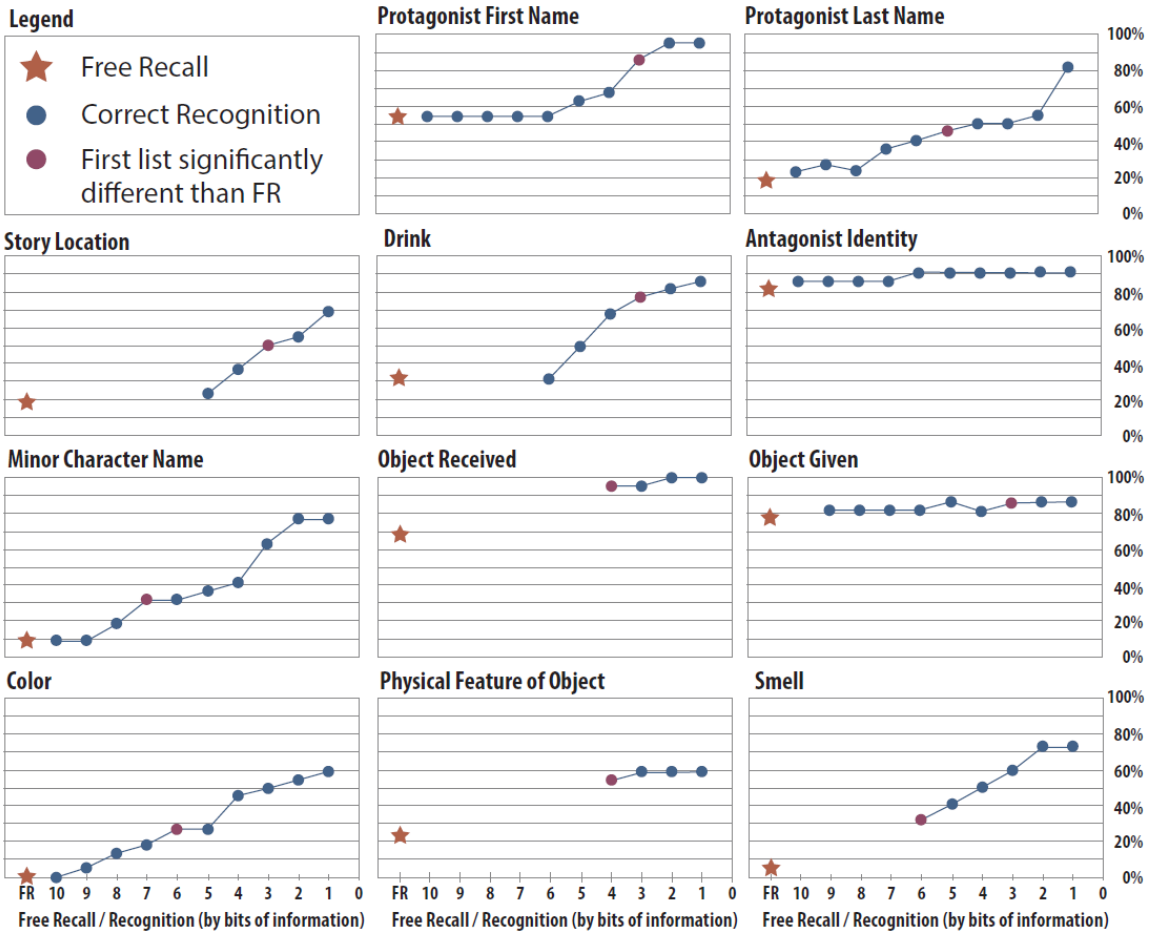


Figure 2.2. Results of McNemar's test, showing free recall performance, recognition performance, and significance.

Performance was also recorded by question as cumulative correct answers at each list size. This gave a steadily increasing measure of number of correct answers for each question for every list size. This data was analyzed using McNemar's Test to find the list size at which performance became significantly different from free recall ($\chi^2 > 3.84$). However, not all questions used every list size and performance on questions was highly variable. One easy question never became significantly different due to a ceiling

effect of performance while others only reached significance at list sizes of 3 or 4 bits (See figure 2.2).

The step-down recognition procedure provided valuable information showing how different questions produce variable patterns of recognition performance. Some questions plateau early and performance never improves. Others show nearly linear improvement over the successive lists. Two interesting trends emerge: 1) Nouns that feature prominently in the story (Protagonist first name, antagonist, object given/received) show high consistent performance. 2) Adjectives and descriptors attached to those nouns show weaker performance. Previous research has already indicated that this performance difference between types of words exists, but it is unclear from these results if that effect is primarily driving these results, or if it is related to frequency of repetition in the story, novelty/uniqueness of the words used, or other properties of the words themselves. Increasing the number of story versions and manipulating those properties directly in subsequent studies will help untangle the relationships between these properties.

2.4 Study Four: Long-Term Memory Effects and Cueing Items

The fourth pilot study of 19 participants also used this step down procedure; however the list sizes were limited to a maximum of 256, because no effects of list size for recognition performance compared to recall were found above that level. Two other variables were also introduced: a second test session was added where participants were asked to perform the recognition test again after an interval of 7-14 days had past

to evaluate long term memory, and half of the participants received a story where the to-be-remembered items were highlighted (by using bold font) so they would act as a cue to draw more attention. This iteration of the study used only the modified random story (reading time 3:52 minutes).

The data were analyzed using a two factor ANOVA with the factors of session and cued/uncued. The results showed that in both sessions participants in the cued condition performed significantly better ($F(1,9) = 340.86$, $p < .001$, $MS = 52.51$) than uncued participants. There was also a significant main effect of session ($F(1,9) = 33.47$, $p < .001$, $MS = 5.16$), indicating higher performance during the second session (See Figure *2.3*).

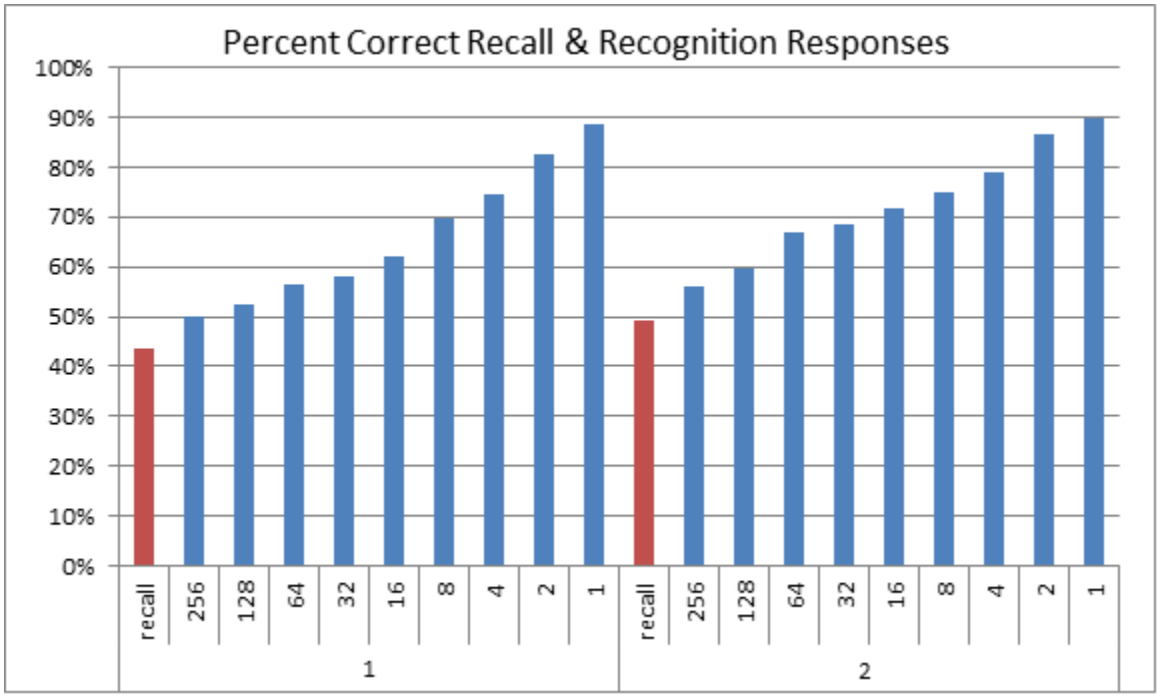


Figure 2.3. Average correct responses (%) for free recall and all list sizes for both conditions during sessions one and two.

A one-way ANOVA was used to analyze the questions and found that performance differed significantly ($F(1,10) = 7.084$, $p = .008$, $MS = 70.306$). Therefore, each question was examined for the list size at which performance on that question became significantly different from free recall performance using a series of T-tests. This was done by comparing average correct answers for each question to the recall average in series, starting with size 2 and working up to size 256 looking for the size at which recognition performance was significantly better than recall performance ($p < .05$) (For an example, see Figure *2.4*).

The bit size of the list where participants were correct with 100% confidence was summed across all 11 questions for each participant. This represented the total number of bits of information they had accurately discriminated between to recognize the correct answers. For session one, the average total bits was 40.47 bits (mean = 3.68 per question, sd = 1.43). For session two, the average was 45.59 bits (mean = 4.14 per question, sd = 1.57). The bit size of the list for the first correct guess each participant made no matter how confident the participant was in that guess was also recorded. This represented the maximum possible performance each participant could have achieved. For session one, average maximum possible performance was 47.21 bits. For session two it was 52.18 bits. The data contained one outlier in the uncued condition who had very poor performance. The analysis was run with and without this participant, but the results were not significantly different.

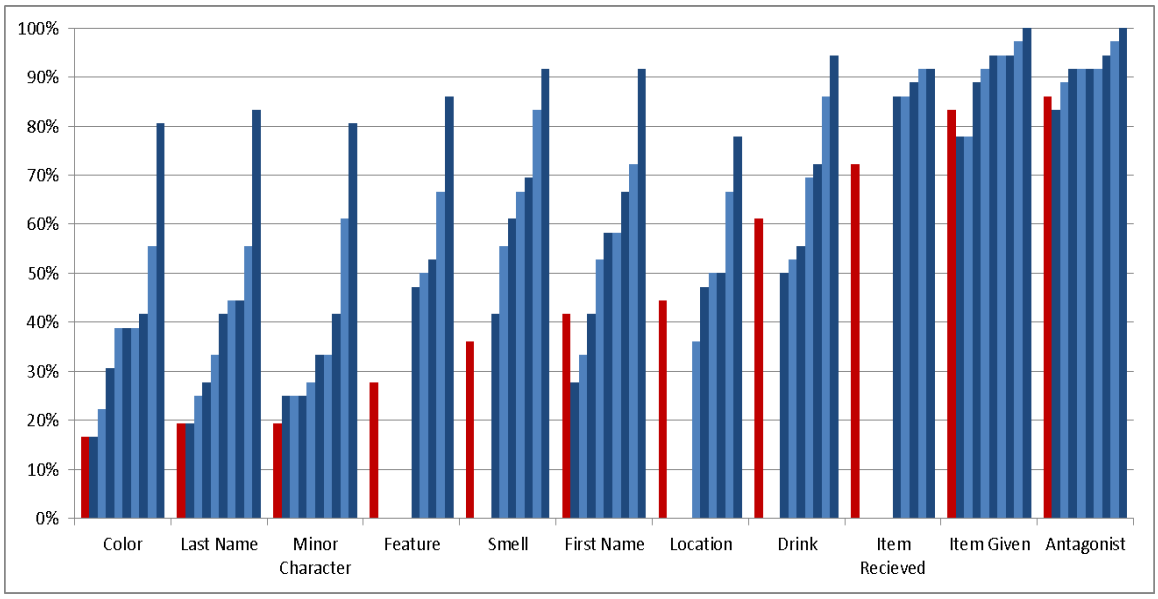


Figure 2.4. Average performance on each question (1-11) at each list size recall (and free recall in red) for cued and uncued participants during both sessions.

2.5 Study Five: Recognition Performance Differences and Names

The final pilot study examined two questions raised by the previous study. First, while performance generally improved with smaller list sizes, these changes were not consistent by list size or question. Therefore, to investigate whether performance was actually related to list size or was due to participants being more comfortable with small lists the final pilot study manipulated list size between subjects using 64, 32, and 16 item lists.

The other primary variable under investigation was performance on names. In the previous study, recall performance for the protagonist's first name was twice that of the last name, or the name of the minor character. This was despite the fact that the participants read the first and last name at the same time. For the new study a variable of combined names was introduced. Half of the participants were asked the protagonist's first and last names as a single question and half were asked those names

as separate questions, to judge how performance changed. In order to maintain a consistent number of questions across all participants, those who were asked the protagonist's name as a single question were asked to produce the first and last name of the minor character as two questions and vice versa.

Other variables under investigation were story version, retention interval, and recall vs recognition. For this study, four versions of the story were used and were manipulated between groups to check for consistency. Retention interval and recall vs recognition were the same as in the previous study. 39 participants took part in the study, with 34 returning for the second session. Two participants were subsequently dropped from the analysis, one for an abnormally long reading time of 12:08 minutes and the other for an abnormally low performance with no correct answers at recall. Average reading time was 4:04 minutes.

The results of the study show three main effects. Two of these effects are the same as in the previous study: participants generally perform better in session two ($F(1,22) = 16.498$, $p = 0.001$) and recognition performance is higher than recall performance ($F(1,22) = 56.285$, $p < 0.001$). There was also a main effect of story version ($F(3,22) = 3.187$, $p = 0.044$). Post hoc analysis showed that this effect was a result of a performance difference between story versions 1 and 4 (story 1 having the lowest overall performance and story 4 having the highest. Mean Difference = 0.1719, $p = 0.010$). The stories used in the final study, based on the same template as story versions 1-4 can be found in Appendix B.

The results for the primary variables did not show an effect of list size ($F(2,22) = 0.957$, $p = 0.400$). Participants tended to do equally well despite the extra 2 bits of information in the 64 list compared to the 16 list. This implies that some of the performance difference in the previous study was in fact due to participants being more comfortable searching through smaller lists, but when forced to use large list they are still capable of the same level of performance. This is an important finding because it implies that the narrative password system could be built with distractor sets of 64, providing 6 bits of entropy per question without sacrificing performance (See Figure *2.5*).

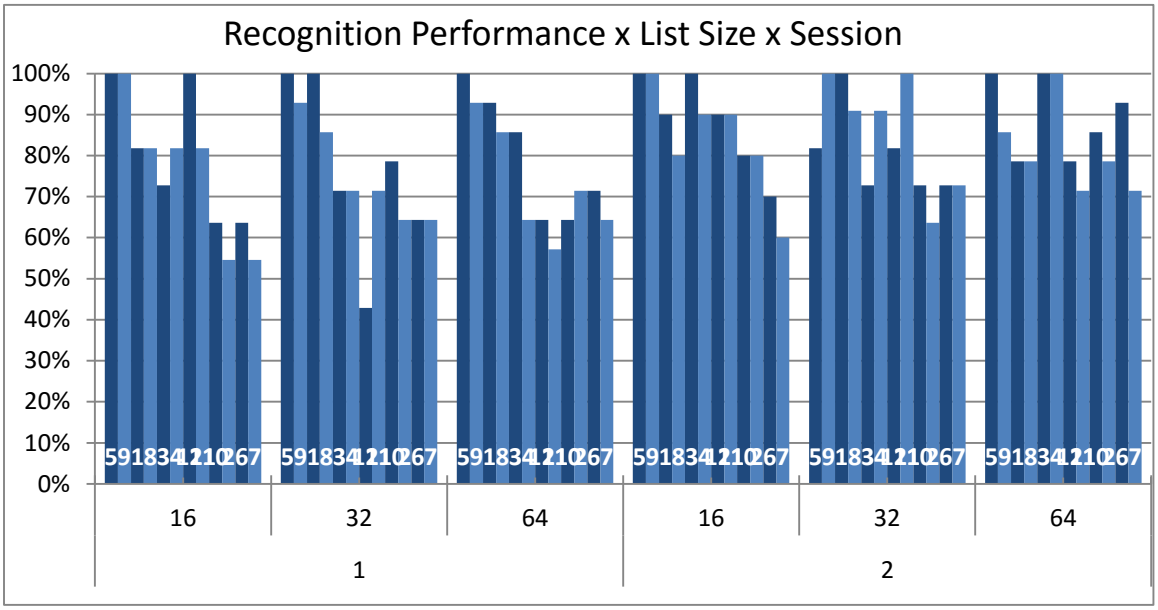


Figure 2.5. Recognition performance on each question by list size during each session.

The other interesting result was performance on individual names compared to combined names. This data was analyzed with a series of pair-wise t-tests that compared performance on first names individually, last names individually, and whole names for both sessions. The analysis shows that while there is no difference between

performance on first names and whole names (t(36) = 0.572, p = 0.571) performance on last names is significantly worse than either(First-Last t(36) = 3.424, p = 0.002. Both-Last t(36) = 3.482, p = 0.001). This result implies that a participant's ability to remember the first name of the character is the limiting factor on their ability to remember the name as a whole. Participants almost never remember the last name without remembering the first name. This has important design implications for the narrative password system as well. If names are going to be used as password characters then it should be the whole name asked as one question or just the first name; last names only drag performance down. (See Figure *2.6*)

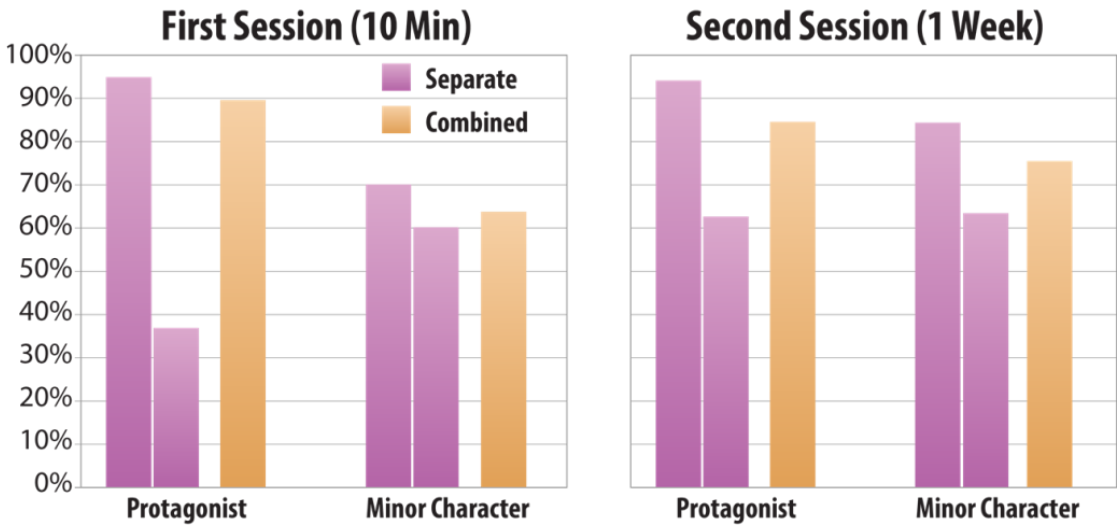


Figure 2.6. Performance on first, last, and whole names by session and question.

Experimental Design

3.1 Comparison of Narrative Passwords to Other Systems

In the final study we compared the narrative password system to two other security systems. Comparison to other systems is a critical variable because it would be

the primary determinant of overall success of the system which had heretofore gone unexamined in the previous studies. We compared narrative passwords to a traditional alphanumeric password of equivalent strength as well as the existing Composite Scene Analysis graphical password system (CSA). CSA and narrative password systems are based on analogous principles but use different modalities, providing a point of comparison on the two most important factors; security and usability. While the narrative password system has higher potential entropy due to the larger number of distractors that can be used, the CSA system showed higher retention rates after the one week interval in previous research (Johnson & Werner, 2008). Understanding the differences between these systems under the same conditions will provide information helpful for future work on both systems.

3.2 New Variables: Multiple Systems, Context, and Retention Interval

Both the narrative password and CSA systems rely on the context provided by either a picture or story to create a usable password. However, our previous studies on narrative passwords have not addressed the degree to which this context helps the participants remember random information with the narrative system. A variation of CSA called Visual Identification Protocol (VIP) looked at the effects of removing the picture context from CSA and found a significant reduction in performance (Johnson & Werner, 2008). Because we believed context to be a key factor that would affect the success of the system, we added a variable to determine how much memory is enhanced by context. Therefore, we included a no context condition in which

participants received both pictures and story objects without the larger context. For the no context CSA system, this was simply a grid of images, the same protocol used in VIP. For the narrative password system this was the list of words and names that make up the narrative password and would normally be connected by the story (a first and last name, drink, location, etc.) which is something we had never previously tested. The authentication procedure was the same; participants chose from the same list which of the items they had seen before in order of presentation.

Apart from the comparison between three different cognitive password systems, there are two other variables we investigated; retention interval and the use of contextualization of the material. The previous studies have only used two retention intervals of five minutes and one week. However, since most passwords are used for much longer periods, to examine the long term effects of system we included longer retention intervals. Therefore, we included testing sessions consisting of a practice session immediately after encoding, a session one week after encoding, and a session three weeks after encoding.

Lastly, as a check on consistency for the systems, we used three versions of the narrative password story (not the same ones used in previous studies) and three different CSA pictures. These were randomly assigned to the participant during enrollment. We also recorded the amount of time participants took when memorizing the different passwords and how long they spent on each of the authentication screens for the narrative password, CSA password, and alphanumeric password. This allowed us

to examine the different enrollment and authentication times of the systems and any potential usability issues related to login time.

3.3 Narrative Password Changes

Based on the previous findings and research, we conducted a study that examined the effectiveness of the narrative password system in an online environment compared to two other password systems. We modified the previous narrative system for use in this setting. First, we modified three of the password questions based on the data from the pilot studies. Minor character's name, color of the object, and main character last name were the three globally worst performing items. For color at least, some investigation provides answers as to why performance was so low. Previously, we cited researcher that showed adjectives are less memorable than nouns (Gentner, 1981; Pavio, 1986; Shepard, 1973). Since color is used as an adjective in the story rather than a noun in its own right that may be limiting participant's memory. Secondly, looking at the incorrect responses for that question reveals that "gold" and "silver" are the two most common incorrect answers. It seems participant's situation model for the story includes the association between the colors gold and silver and the object focus of the story. When participants have difficultly remembering the correct answer, they rely on this association to provide it, and since the correct answer violates that assumption they answer incorrectly. Since it is the lowest performing item it was dropped from the final test.

Minor character name and last name may perform poorly because memory for names tends to be low and they are not centrally related to the plot. Therefore we added a descriptor to the minor character name, unique job titles and professions to modify the name and asked the protagonist's full name in the first question. Combined first and last names were shown to have equivalent performance to first names and using both allows the last name to potentially act as a cue, albeit probably infrequently. To further increase the saliency of the names now that both names no longer have to be randomly selected, we chose more novel names with a lower frequency of use. We also made first and last name alliterative to increase their discriminability. We used authentication lists of 16 items and 9 questions as the password, which will still provide 36 bits of entropy; the equivalent to entering a 6 character alphanumeric password including numbers and uppercase and lowercase letters. This allowed us to both match the strength of the narrative system to the CSA system and choose the 16 most discriminable items for each question (Johnson & Werner, 2006, 2007, 2008). These options were chosen based on a lack of conceptual overlap with other items, novelty in language, and comprehensibility within the story.

Methods

The study was a 2 x 3 x 2 mixed factorial design and took place online over the course of three weeks with a total of 103 participants, 86 of whom returned for session two, and 83 returned for session three, six of those who returned for session three had not participated in session two. 28 of the participants were male, 75 were female, and

the average age was 21.6 years (SD = 10.8). Six participants indicated that English was not their native language and five reported they had or may have had a condition that could inhibit memory, however these participants did not appear as outliers so they were included in the analysis. The specific variables were system context (Present vs. not present, between), password type (Alphanumeric vs. Narrative vs. CSA, within), and retention interval (1 week and 3 weeks, within). Participants were recruited through the University of Idaho SONA Systems site and compensated with class credit. The study itself was built using Cognilab, an online experiment creation program which randomly assigned participants into their conditions, controlled stimulus presentation, and collected performance data (Cognilab, 2015).

Participants signed up for the study on the SONA Systems website and received a randomly assigned link to one of three versions of the study. Each version used a different randomly generated alphanumeric, narrative, and CSA password. Then participants read and agreed to an online consent form, entered their demographic information which was used to match participants across different sessions, and were assigned to either the system context or no system context condition (which was only relevant in the enrollment session). The order of passwords was held constant due to constraints imposed by the CogniLab software used. Participants first received their randomly generated alphanumeric password which appeared on screen for at least 60 seconds, during which they viewed and memorized the sequence. After 60 seconds the participant had the option to skip forward to the next section, if they did not skip forward the password remained visible for a total of 240 seconds before it disappeared.

Next the participants were shown their assigned CSA picture which they could view for up to 240 seconds, with the option to move to the next section available after 60 seconds. In the non-context condition this consisted of a grid of nine images that remained static on the screen. In the context condition, each of the nine image elements flashed on the screen for three seconds before being replaced by the next image element to familiarize the participant with each component of the image ensuring they had seen all of them. After all of the image elements were seen individually, the full composite image was displayed for the remainder of the enrollment time (minimum of 60 seconds, maximum of 240 seconds). Finally, the participant received their randomized short story or word grid. The story was broken into six paragraphs that were presented on separate screens. On each screen the participants were allowed to read for as long as they needed with the button for proceeding to the next page appearing after 15 seconds. This was intended to prevent participants from immediately skipping to the next page. In the non-context condition the participants were shown a grid of words, with an option to continue appearing a minimum of 60 seconds later. Similarly to the other conditions, the grid disappeared after 240 seconds, indicating to the participant it was time to move on. For each enrollment screen the time which participants took to enroll was recorded. Participants then continued to the first authentication section.

The authentication section asked participants to input each of the three passwords they had seen in the same order as enrollment. First, a text box appeared which prompted the participant to enter the alphanumeric string they had seen before.

If the input provided was less than five characters or greater than seven the participant was prompted to restrict input to a range between 5-7 characters. This was designed to avoid unintentional entry errors (e.g., using the return key too quickly or entering a string of characters that was inappropriately long). Participants were informed that they had three attempts to enter the correct string and that the passwords were case-sensitive. After each failed attempt participants were reminded of the number of attempts remaining and the parameters of the test. After a successful entry or three failures the participants proceeded to the next section. Reaction times, text entered, and number of attempts were recorded.

During the CSA authentication system participants were shown nine grids of 16 image elements and asked to identify which image element they had been shown previously. Again, the participants had three attempts to select all the correct answers. If the participant made an incorrect selection they were informed at the end of the attempt that one or more of their choices was incorrect and asked to try again. They were also informed of the number of attempts remaining. They were not told which of the images had been the incorrect choice.

After a successful trial or three incorrect attempts participants continued to the narrative password authentication section. The procedure for this section was the same as the CSA section, with participants choosing the words they had previously viewed from nine lists of 16 items. If participants made a mistake they were asked to try again and not given specific feedback about which items were incorrect. In both conditions

the program recorded the time to make each selection, which image/word was chosen, whether that choice was correct, and how many attempts were made. After a successful trial or three failures participants were asked for feedback about the experiment and reminded they would receive an email in one week for the next session.

Approximately one week later the participant were sent an email containing a link that allowed them to access the system again for the next part of the experiment. The participants were asked to take the test within three days of receiving the email. 90% of participants responded within this window, with all who responded coming back within 8 days of receiving the email. The experimental session was exactly the same as the previous test, minus the enrollment phase. To verify which participant was completing the study they were asked to enter their SONA ID number, email address and demographic information at the beginning of the test again. Participants were asked to enter or select their alphanumeric, CSA, and narrative passwords in that order. They were given three attempts to enter each correctly and given feedback about whether or not they had correctly entered the password. The authentication time, text entered, images/words selected, and number of attempts were recorded again. Participants were allowed to enter feedback again as well. The participants repeated this procedure again at three weeks after enrollment, along with a brief survey about the experiment.

Due to a problem with three of the links that were sent out for sessions two and three, 19 participants in session two and 19 in session three initially participated in the

wrong version of the test. This problem was identified within two to three days of participants receiving the incorrect link and they were sent the correct condition. Of the 19 who received and followed bad links, 11 returned and participated in the correct version of the test during session two and 14 returned and took the correct version in session three. The performance of these participants was not distinguishable from participants who received the correct link and they were left in the study.

4.1 Materials

The primary materials used in this study were the nine passwords, including the authentication screens used for each of the three types, as well as the Cognilab (2015) software which is currently in beta testing. The three alphanumeric passwords were created from randomly selected numbers and letters as follows: Version one was “jCM4NG”. Version two was “4nPCXG”. Version three was “aeY9BY”. Unlike in previous studies, these alpha numeric passwords were case sensitive, allowing them to more closely match the bit value of the CSA and narrative passwords. A length of six characters and a pool of 62 produces a bit value of 35.7, compared to 36 for the other two systems.

The three CSA password images were the same used in Johnson & Werner (2013) and were created with Adobe Photoshop using stock images. All three images can be found in Appendix A. The story passwords and elements used were new versions of the same story used in the pilot study. After careful selection of the pool of 16 options three new stories were made by randomly choosing from those pools. All three

stories can also be found in Appendix B. The authentication screens with the distractor items used in the study can be found in Appendix C.

Figure 4.1. Three different image elements used in the CSA system. See Appendix A for complete images.

Elizabeth English had just poured herself a **glass of coffee** and flipped open a book she'd been looking forward to reading when a sudden whooshing sound made her look up in time to witness a **Fairy** materializing in the **library**. For a moment **Elizabeth** considered running out of the **library** and then down the road to the church to fetch **Father Ingram** – but she soon dismissed the idea. Her familiarity with otherworldly creatures such as this **Fairy** was poor indeed, but she realized that the **Fairy** might very well take offence at **Father Ingram** bursting in. Under no circumstances would she have her home turned into a battlefield. Furthermore, **Elizabeth** was not sure that **Father Ingram** had any more experience than she did. Was calling a **Father** the best procedure for dealing with a **Fairy**? **Elizabeth** decided that trying to figure out the answer was taking too much time, so instead she cleared her throat and simply said: "Good afternoon!"

Figure 4.2. The first paragraph (1 of 6) of version one of the authentication story. See Appendix B for complete stories.

Results

This study had three key hypotheses: First, we assumed that Narrative passwords and CSA passwords would perform better than alphanumeric passwords for longer retention intervals. Second, overall performance would degrade over the retention intervals in all conditions, but would degrade faster for alphanumeric

passwords than for narrative or CSA passwords in the second and third sessions. Third, participants who received passwords within a context (either as a scene consisting of the different image elements or the short story containing the verbal password elements) should perform better than those who do not. We examined enrollment and authentication/login times and attempts to assess the usability of the different systems and determine if there was a difference between the password versions. We will look at the results of each of these hypotheses in turn.

5.1 Login Success Rate

During the initial enrollment phase 77.7% of the participants were able to successfully authenticate within three attempts with their assigned alphanumeric password. This percentage was 98.1% with the CSA password, however it dropped to 69.9% with the narrative password. This difference is statistically significant with the CSA password performing better than the other two passwords according to a non-parametric repeated sample Cochran's Q test of successful/unsuccessful login attempts ($Q = 29.911$, $p < 0.001$). The variables tested with this procedure were either the same password type by the three retention intervals or each password type within the same session. (See Figure *5.1*)

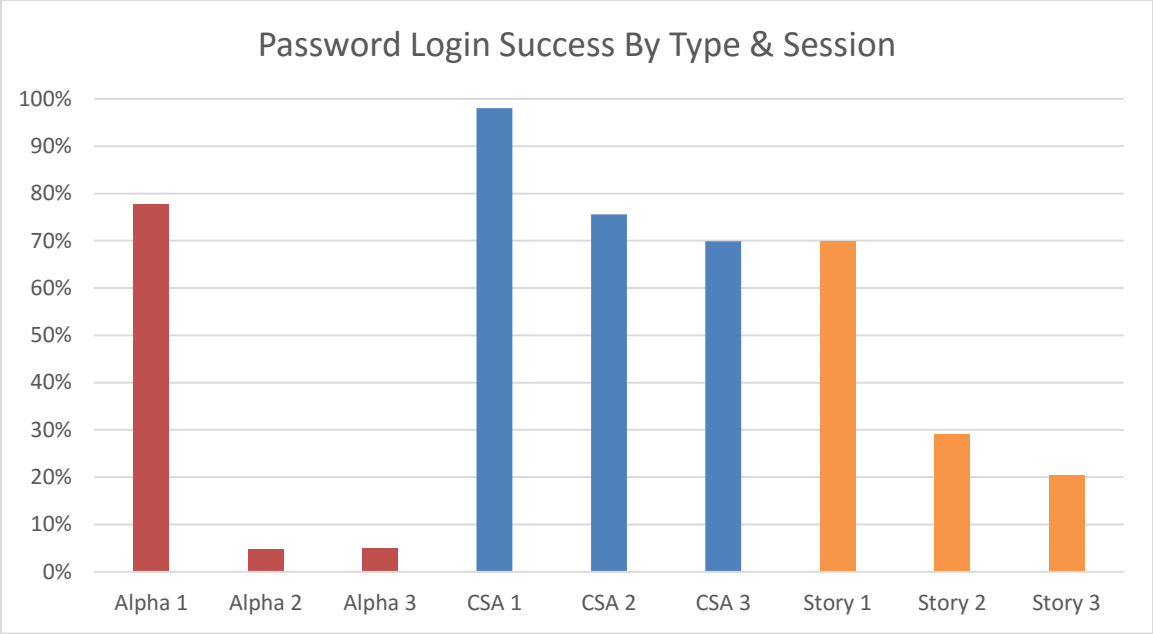


Figure 5.1. Login success by password type and session. 103 participants in session 1, 86 in session 2, and 83 in session 3.

During session two the percentage of successful logins for each password type did drop. Alphanumeric passwords saw the largest decrease to 4.65% of participants successfully logging in after one week, a drop of 73.0% (Session 2 vs. Session 1, $Q(2) = 0.753$, $p < 0.001$). CSA enrollment dropped to 75.6% successful logins, a change of 22.5% (Session 2 vs. Session 1, $Q = .219$, $p < 0.001$). Finally, narrative password login successes dropped to 29.1%, a 40.8% decrease over the course of the week (Session 2 vs. Session 1, $Q = 0.466$, $p < 0.001$). Overall these differences remained significantly different, $Q = 91.460$, $p < 0.001$.

The login rate for alphanumeric passwords did not change from session two to session three over the two week interval, this is most likely due to a floor effect as only a total of four participants were able to successfully login in either session. CSA authentication rate dropped from 75.6% to 69.9%% over two weeks; a change which

was not significant (Session 2 vs. Session 3, Q = 0.750, p = 0.388). Narrative password authentication dropped 9.0% to 20.4%, which was a larger change than CSA, but still not enough for significance (Session 2 vs. Session 3, Q = .642, p = 0.424). Overall session differences remained Q =82.172, p < 0.001.

5.2 Effects of Context

45 participants received passwords without context provided and 58 received a contextual password. The table below summarizes the login success rate for each of the three password types by context and session.

	Session 1		Session 2		Session 3	
PASSWORD TYPE	**Context**	**No Context**	**Context**	**No context**	**Context**	**No Context**
ALPHANUMERIC	79.3%	75.5%	4.2%	5.2%	6.5%	2.7%
CSA	100%	95.5%	79.2%	71.1%	80.4%*	56.8%*
NARRATIVE	65.5%	75.5%	20.8%	39.5%	23.9%	16.2%

Table 5.1. Login successes by context, session, and password type. * indicates a significant result.

The presence or absence of context did not have a significant effect on success rates for any password type or session, with one exception for CSA passwords in session three. A series of Chi Square tests were run to determine if test version or context had an effect on login success rate. Context was significant for CSA during session 3 when those in the context condition maintained a higher login rate than those in the no context condition (X^2 (1, N = 83) = 5.462, p = 0.019) (See Table 5.1). Test type was also significant at one session and type only. During session one, the version 1 alphanumeric password performed significantly better than the other two versions (V1 = 12 successes, V2 = 8 successes, and V3 = 3 successes) (X^2 (2, N = 103) = 7.143, p = 0.028). This is not a

problematic result because test version was added as a consistency check on the Narrative and CSA passwords which did not show any effects and it is not unexpected that certain randomly chosen letters and numbers might be more memorable than others.

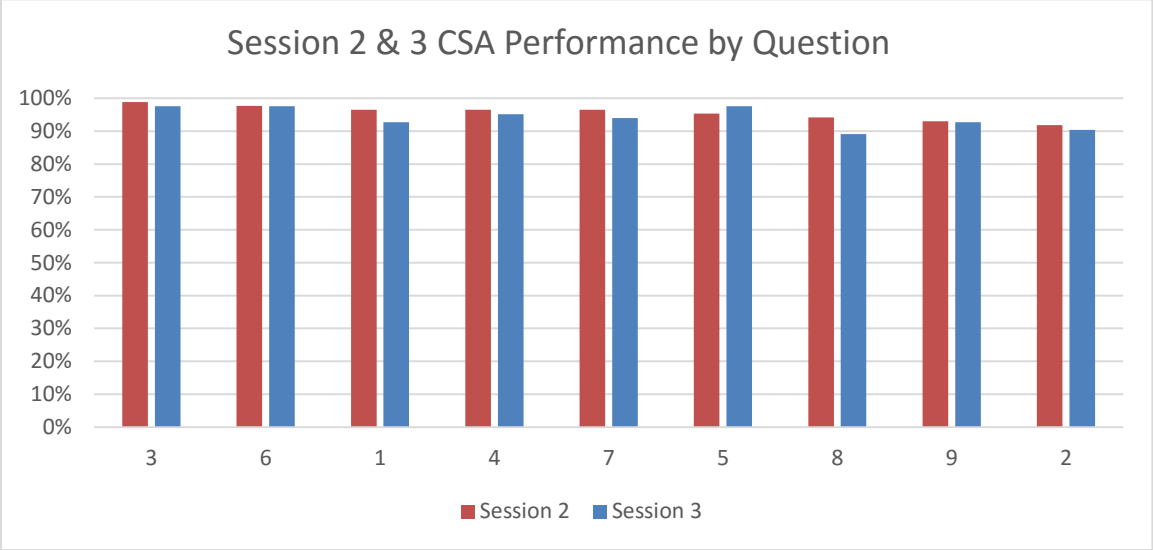
5.3 Generalized Linear Mixed Model Binomial Logistic Regression

Since our primary measure of success was a binary outcome we could not analyze our full model with a MANOVA. Instead we used SPSS to run a generalized linear mixed model (GLMM) binomial logistic regression. This procedure uses a regression model to predict target binomial data (login success) using factorial conditions (password type, session, and context). We fully analyzed all main effects, 2-way, and 3-way interactions. Since the model is more comprehensive and sensitive than the standard MANOVA model, we set a more conservative p value of 0.001. We found significant main effects for session ($F(2,182) = 84.863$, $p < 0.001$) and password type ($F(2,154) = 92.498$, $p < 0.001$). Context was not significant at the same level. We also found significant interactions of context and session ($F(2,182) = 7.206$, $p = 0.001$), context and password type ($F(2,154) = 16.518$, $p < 0.001$), session and password type ($F(4,182) = 11.418$, $p < 0.001$), and a three way interaction of all the variables ($F(4,182) = 6.191$, $p < 0.001$).

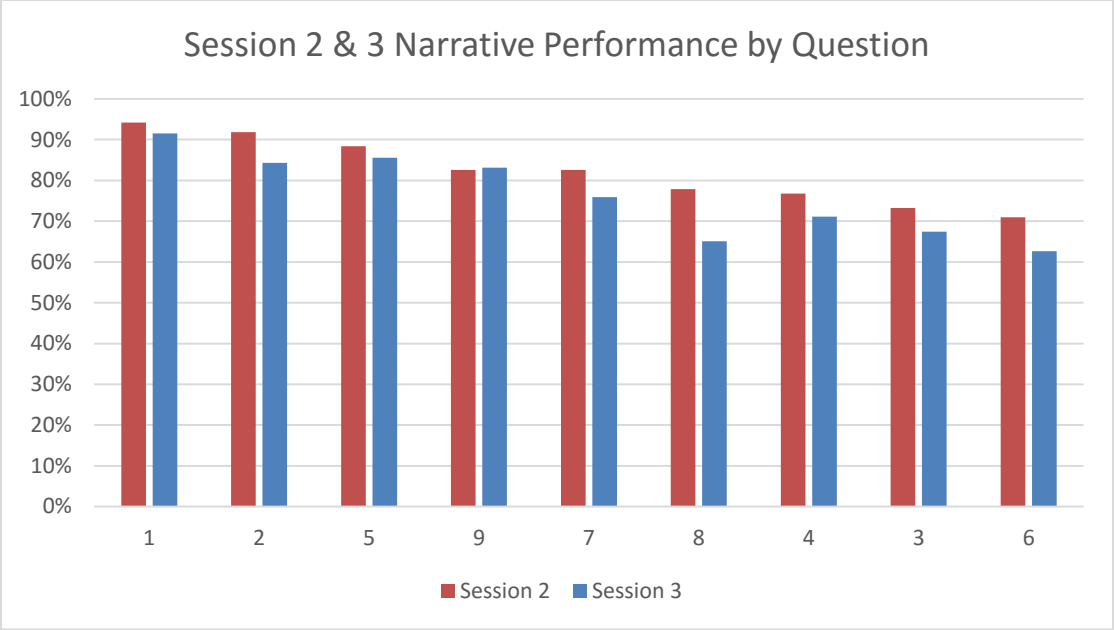
5.4 Question Analysis

An analysis of the individual questions in session two of the study reveals that there do not seem to be “barrier” questions with markedly lower performance that

hinder a majority of participants (See Figures *5.2 & 5.3*). The two lowest performing story questions, 3 & 6 ("Drink" and "Item Given") still have success rates above 70% in session 2 and above 60% in session 3, though this performance is notably lower than in the final pilot study for these items. This may indicate a problem of miss-correction, where participants do not know which question prevented them from accessing the system so they change an inappropriate item.



Figures 5.2. Performance on each of the 9 questions for CSA, created from each participant's best attempt.



Figures 5.3. Performance on each of the 9 questions for Narrative passwords, created from each participant's best attempt.

Another way to examine how individual questions impacted Narrative Password performance is to compare performance of each question to the number of repetitions within the story. We documented in the introduction that one factor that influences memory for words is how frequently that word is repeated within the text that is being memorized. To analyze this we took average performance for each question in the Narrative Password and correlated that with the number of times the answer appears in the text, broken down by context and session. (See Figure *5.4*)

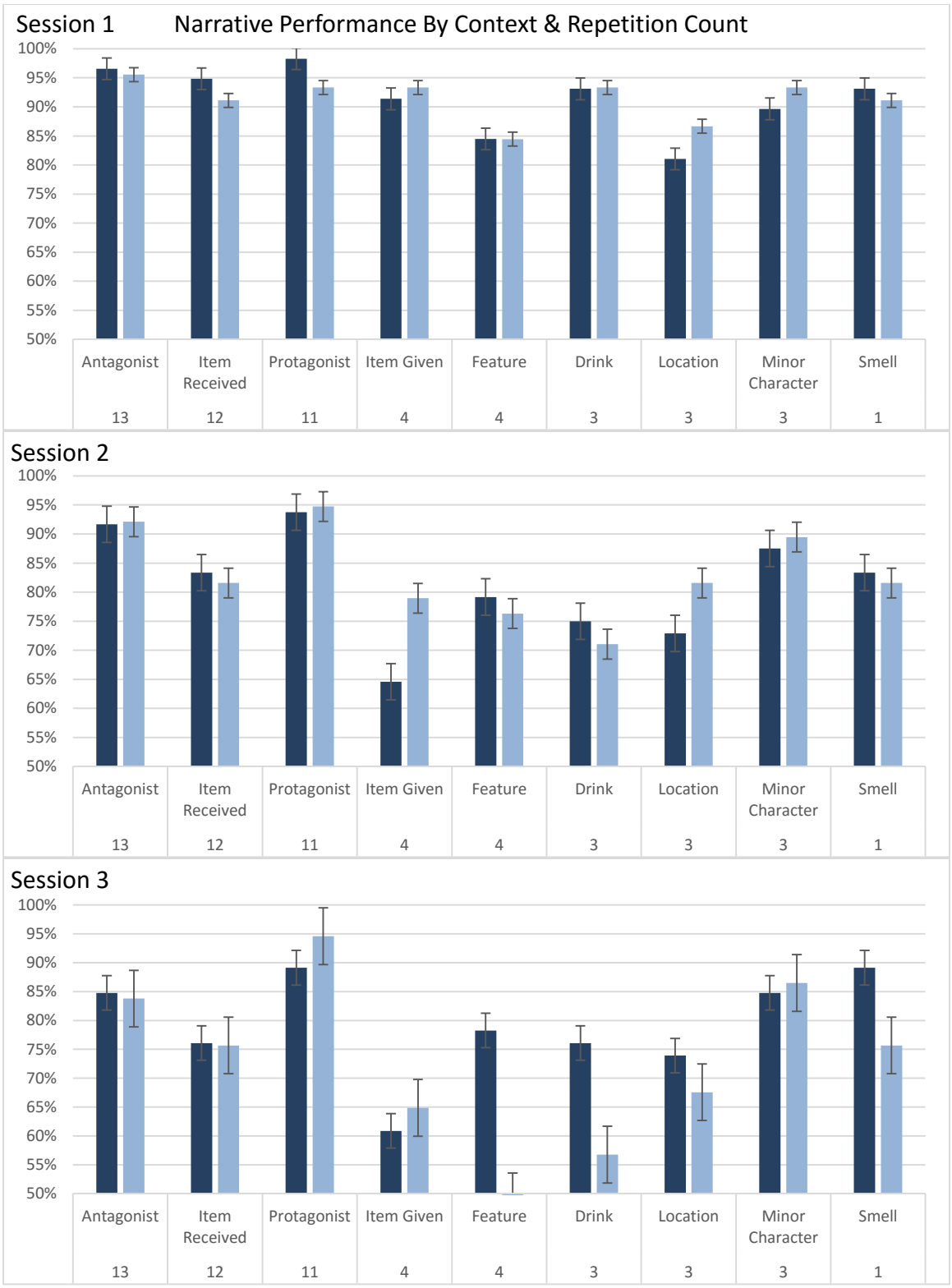


Figure 5.4. Average performace on each question by the content of the question and number of repetitions within the story. Dark blue bars represent the context condition light blue is no context Error bars relfect standard error. (S1 n = 103, Pearson's r = 0.52)(S2 n = 86, Pearson's r = 0.57)(S3 n = 83, Pearson's r = 0.37)

5.5 Enrollment and Login Time

Participants spent different amounts of time learning each of the three password types, though they were only required to spend at least one minute on each or 90 seconds on the story (15 second per each of the six pages). Participants averaged 1:09 minutes on the alphanumeric password, 1:18 minutes on the CSA password, and 2:54 minutes on the narrative password. An ANOVA performed for the different password types showed context did have a significant effect on enrollment time, ($F(2,97) = 8.370$, $p = 0.005$) and ($F(2,97) = 39.019$, $p < 0.001$) respectively. In the no context condition enrollment times for the story and pictures are approximately the same: 1:37 minutes for the pictures and 1:40 minutes for the story. However, in the context condition the enrollment time for the pictures drops to 1:03 minutes (Also the highest performing group in session 3) and the reading time for the story increases to 3:52 minutes. There was no significant difference in alphanumeric enrollment times, nor was there any difference in the different test versions.

Additionally, reaction times for password input were recorded for purposes of usability analysis of the password varieties. This was done by looking at the input times from the successful attempts to input the three password types. The chart below breaks down successful input time by password type, session, and context (See Figure *5.4*). Alphanumeric entry times were the fastest overall, followed by CSA, then narrative. Alphanumeric and CSA entry times rose during session two, while narrative entry time dropped. An ANOVA of these times showed context and test version did not have a

significant effect on login time, but password type had a significant effect ($F(2,194) = 10.016$, $p < 0.001$) as did session ($F(2,194) = 71.055$, $p < 0.001$) and the interaction of the two ($F(4,388) = 4.008$, $p = 0.003$).

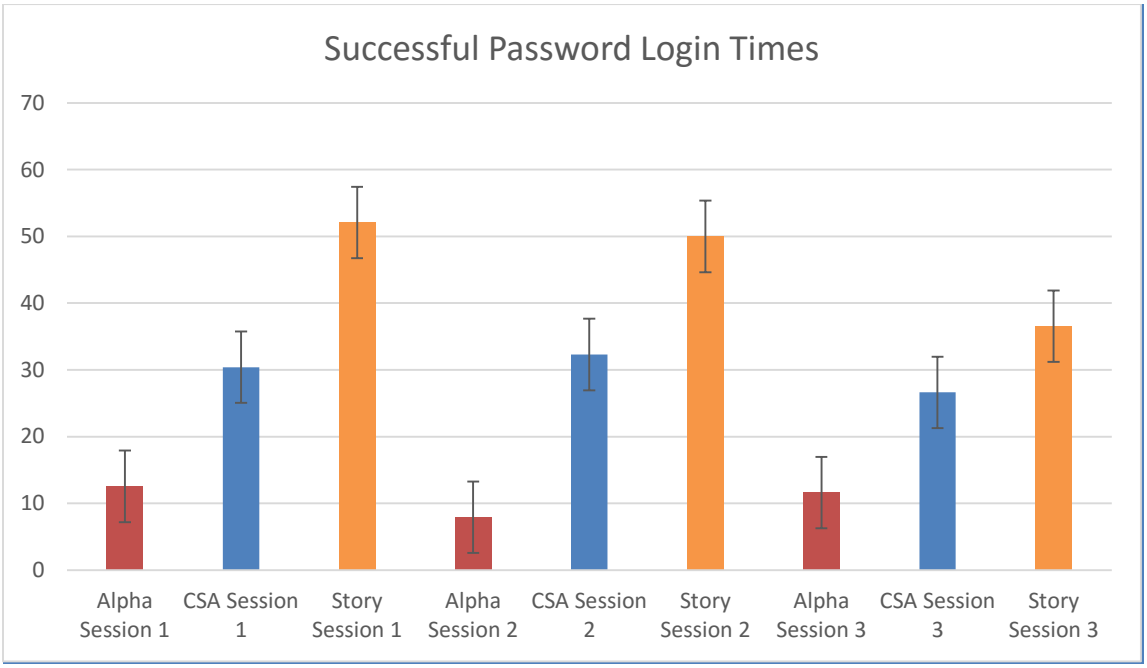


Figure 5.5. Average login times by password type and session. All times are in seconds and comprise the participant's successful attempt. Bars reflect standard error of the mean.

To better understand how long term practice might influence login times we restricted the range of the data to the ten fastest successful login attempts for each password type over each session. These represent an estimate of approximately how fast a consistent user could authenticate using each system. Alphanumeric login times do not change for sessions two and three because there were fewer than ten successful logins. What we can see in Figure 5.8 is that the login times drop by an average of 11 seconds, CSA login times for each session drop by an average of 14 seconds to an average of 16 seconds, and Narrative Password login times drop by 18 seconds to an average of 29 seconds.

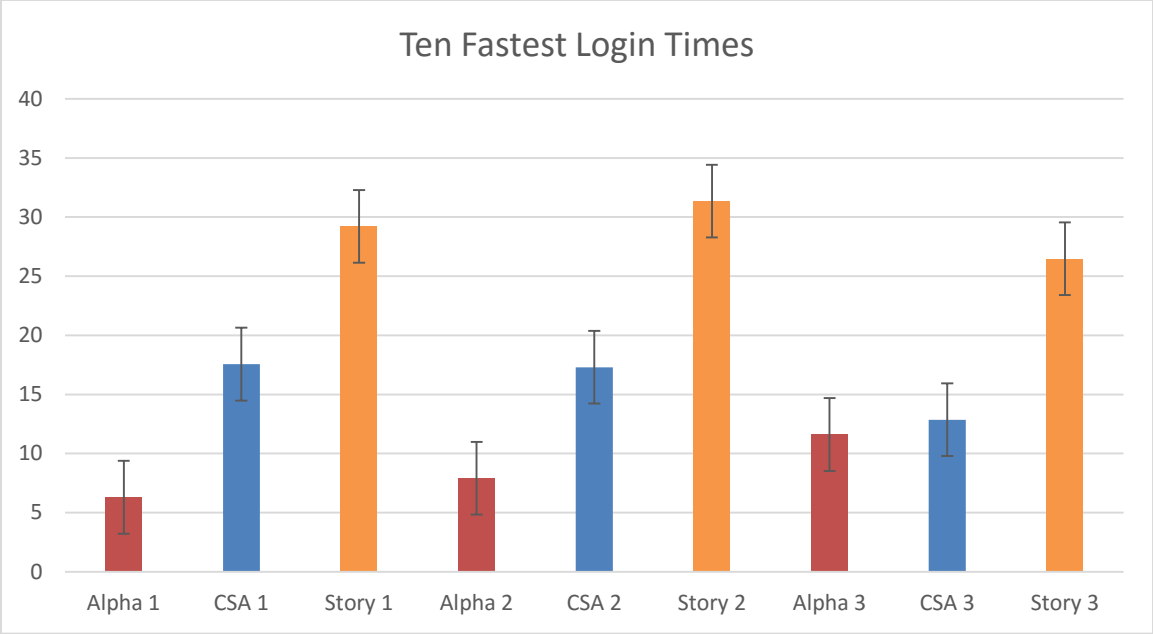


Figure 5.6. Average of the ten fastest successful authentication times for each password type across session. All times are in seconds, bars represent standard error of the mean.

5.6 Usability Survey

After completing the third session of the experiment, the participants received a short survey about their experiences with the different password systems. They were asked to rate the difficulty of each system on a 5 point Likert scale (Very Easy-Very Difficult), choose which password system they felt was easiest to remember, choose which system they would prefer to use for accessing a website they used infrequently, and answer two open-ended questions about why they preferred certain systems over others and their general experience with the experiment. Due to an error with the link sent to one group of participants, 19 participants were given a link to the incorrect test version, however these participants were sent a correct link later and 11 retook the test. The repeated answers were dropped from the analysis and we only used answers from the correct version. Interestingly, since the emails used to provide links were those

provided in the first session, participants who missed session two returned for session three. Their results were dropped from the ANOVAs, but their survey responses were recorded, resulting in 103 survey responses. The surveys were sent out to all participants, regardless of attrition, which resulted in participants responding even if that had not completed all three sessions. These results were included.

Participants ranked Alphanumeric passwords the most difficult (Mean = 4.3), Narrative Passwords as the second most difficult (Mean = 3.1) and CSA as the easiest (Mean = 1.7). When asked which system they would prefer to use to access infrequently used information participants primarily responded that they prefer CSA. 10 participants responded that they would choose alphanumeric, and 5 chose narrative. Results were similar when asked which system was the easiest to remember; 3 participants chose alphanumeric and 4 narrative. (See Figure *5.7*)

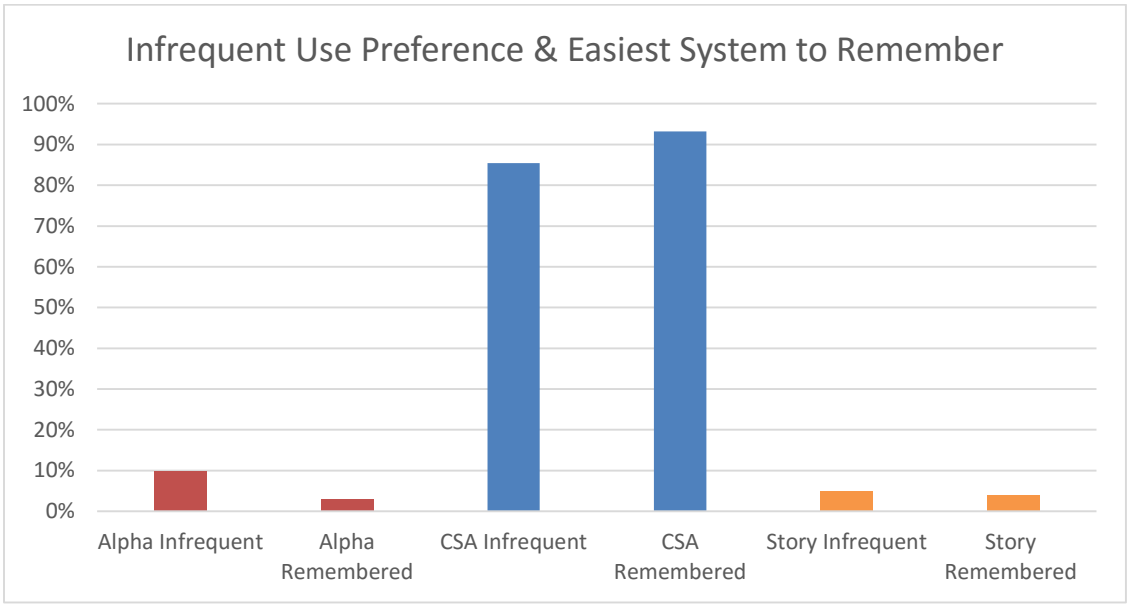


Figure 5.7. Percentage of participants who chose different password types for access to infrequently used data and password type that was easiest to remember.

Discussion

In this study we sought to test the Narrative Password system that had been developed through pilot testing against two other passwords systems on the basis of ability to successfully enroll/authenticate and retention over a course of weeks. We also examined the fundamental question of whether or not context plays the significant role we expected and participant's subjective feelings on the different systems. There are several important results that merit discussion. First, although all of the different password types show a drop in performance over the course of the study, CSA passwords are the most reliable and contrary to our hypothesis perform far above narrative passwords. Second, we found no overall effect of context on the participant's ability to recognize the correct password characters, which at first glance seems contrary to what we expected to find and contradicts previous research (Johnson & Werner, 2008). However, as in Johnson and Werner (2008) CSA passwords showed an increasing benefit of context with increasing retention interval – unlike the other passwords systems we tested. Third, there were no individual questions for narrative passwords with performance below 60% correct. This implies that no one question imposed a serious barrier to success but that some factor of the system at large prevents successful use. Forth, password enrollment time are the shortest for the highest performing password type, and login time was shorter in session 3 than session 2. Finally, the survey responses from the participants indicate a clear preference for CSA passwords.

The data collected in this study strongly indicates that narrative passwords do not hold up compared to CSA passwords in a large scale, realistic setting. During the initial enrollment session the narrative passwords were only retained at approximately the same level as the alphanumeric passwords, and at a 20-25% lower rate than CSA. Furthermore the performance on narrative passwords decayed faster than CSA over the first week of the experiment, though not as much as the alphanumeric. The extremely low alphanumeric performance does provide strong evidence of participants following the rules of the experiment; the alphanumeric password would have been the easiest to cheat on by writing the answers down, but that does not appear to have happened. If participants did not cheat on that portion of the test it seems unlikely that they would have cheated otherwise. The most likely cause of the low performance on the test can be summarized by the feedback provided by one participant:

> "I did not even think of these things for long after I did the last test. I believe it should be made clear at the beginning to remember these things. I only can remember things I do in a weekly to daily activities fashion…"

Though the purpose and participant's responsibilities were clearly stated, it is possible that not all of the participants took the experiment as seriously as they would have when trying to memorize a password that they needed to use in day-to-day life. In future studies it may be prudent to add more incentive to complete the login process, like requiring participants to actually access data with their passwords. Even if

that is the case however, the fact that the CSA system outperforms the others by a significant margin only adds to the conclusion that it is the better system.

CSA maintained very high performance after three weeks, with only 30% of participants having lost the ability to remember a random password. 50% of participants lost the ability to access the narrative password over the course of the experiment. Alphanumeric performed the worst overall; only 5% of participants could remember their password after just one week. This is much lower than in the previous studies conducted by Johnson and Werner (2008). One explanation for this difference is that the passwords used in this study were case-sensitive unlike the previous studies. When allowing for case-insensitivity we see the number of successful password entries in sessions two and three double to 10% (a total of 8 participants).

The most surprising finding of the study was that context did not have a main effect for login successes, even with alphanumeric passwords excluded from the analysis. There are a couple of potential explanations for this lack of effect that can be explored. The first explanation is that again, participants were not paying close enough attention for the presence of context to have a large effect. Secondly, context may have a smaller effect size than we previously thought and a larger study is required to demonstrate significance; we do see an effect for context but only in session three, and we do have the interaction effect with context and session. Thirdly, there is a potential performance ceiling effect for CSA and that context provides a different benefit, such as better long term retention or faster memorization time. This

is evidenced by the drop in enrollment time for CSA passwords from the non-context to the context condition. Having a composite image may allow users to memorize the password more efficiently. These effects could be more thoroughly investigate in a study that focused on CSA passwords more closely.

From a usability perspective, some work still needs to be done to make the cognitive passwords we tested functional enough for public use. Password enrollment time is a factor that will significantly influence how willing a participant is to use a particular system. It is unclear what the average time is for a participant to create a password themselves, it depends on a number of factors such as reusing an old password, or trying to find one that conforms to the system's requirements. Four minutes of reading a story may be more than participants are willing to do, but one minute of looking at pictures may be acceptable. Participants were required to look at the password for a shorter amount of time in this study than previously, which may account for some of the difference. Finally, we could not find any studies that have examined participants' willingness to use randomly generated cognitive passwords compared to self-generated passwords.

Furthermore, the login time for CSA and narrative passwords are longer than participants maybe be willing to tolerate for frequent use. 30-60 seconds per attempt to login to a system that requires frequent use (say a smartphone) is definitively too long, although these times would likely decrease with practice but further studies will be required to determine how much of a decrease. Our approximation of continued

practice by looking at the ten fastest login times shows that narrative passwords still only decrease to 26 seconds whereas CSA drops all the way to 13 seconds, or half the time of a narrative password with the same level of security. However, 25-60 seconds is probably shorter than the time required to reset a password on a system that only requires infrequent access (certain websites or perhaps financial information). This is assuming the cognitive password is stable enough over that period to not require a reset, but the evidence seems to be pointing in that direction. We do also see a significant drop in login time from session two to session three. Closer examination of the data shows that this is for two reasons. Participants who successfully logged in during a previous session had already practiced their password and could enter it more quickly in session three, resulting in shorter login times. Participants who had previously been unable to remember their password often simply gave up and clicked through their three login attempts as quickly as possible without trying. Future studies could examine how much login time decreases with different input strategies and different amounts of practice to see if our estimations using the ten fastest times are accurate.

User acceptance seems to be predominantly in favor of CSA passwords. Participants overwhelmingly chose it as their preferred system and rated it as easy to use compared to the other systems. More participants said that they would prefer to use alphanumeric passwords to access an infrequently used system than narrative passwords despite their lower performance. In the qualitative response most of these

participants cited their familiarity with alphanumeric passwords as the reason for this choice.

One finding from the qualitative data that does provide a glimmer of hope for narrative passwords is that 17 of the participants, while explaining what strategies they used to remember the passwords, stated that they created their own story to remember the CSA password. This might not have been a story in the same sense as the narrative password, but it provides a line of inquiry for future work with narrative passwords. It may be possible to find a compromise between purely random stories and user generated passwords. Perhaps a system could provide randomly generated items, names, or events and ask the participant to link them together into a story. It could as be included in the instructions for CSA passwords; suggesting users think of a story to improve retention of their password, in a sense combining the two systems.

In conclusion, while evidence does not seem to support the idea that narrative passwords are a viable alternative to traditional passwords, composite scene analysis does appear to show promise. Future experiments are currently under development to speed up the login times of CSA passwords and could incorporate results from this study to encourage more focused participation from experimental subjects and further investigate the role that context plays in aiding memory for randomized graphical material.

Works Cited

Adams, A. Sasse, M.A. & Lunt, P. (1997). Making passwords secure and usable. In Thimbleby, H., O'Connail, B., & Thomas, P. (eds), People and Computers XII: Proceedings of HCI 1997, 1-19.

Andriotis, P., Tryfonas T., Oikonomou, G., & Yildiz, C. (2013.) A pilot study on the security of pattern screen-lock methods and soft side channel attacks. *Proceedings of the sixth ACM conference on Security and privacy in wireless and mobile networks, pp. 1-6.* New York, NY: ACM.

Baddeley, A. (2000). The Episodic Buffer: A New Component of Working-Memory? *Trends in Cognitive Sciences*, 417-423.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology.* Cambridge: Cambridge University Press.

Biddle, R. Chiasson, S. & van Oorschot, P.C. (2011). Graphical Passwords: Learning from the First Twelve Years. ACM Computing Surveys 44(4)

Brainard, J., Juels, A., Rivest, R.L., Szydlo, M., & Yung, M. (2006). Fourth-factor Authentication: Somebody you Know. In *Proceedings of the 13th ACM conference on Computer and Communications Security, pp. 168-178*. New York, NY: ACM.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence Memory: A Constructive Versus Interpritive Approach. *Cognitive Psychology*, 193-209.

Brostoff, S. & Sasse, M.A. (2000*). Are passfaces more usable than passwords? A field trial investigation.* People and Computers XIV—Usability or Else!, 405-424.

Brown, Alan Bracken, Elisabeth Zoccoli, Sandy & Douglas, King (2004). Generating and Remembering Passwords. *Applied Cognitive Psychology* 18, 641-651.

Burr, W. E., Dodson, D. F., & Polk, W. T. (2006). *Electronic Authentication Guideline.* NIST Special Publication.

Cognilab Corperation (2015). *Cognilab web access.* Retrieved April, 2015 from http://www.cognilab.com/

Cubrilovic, N. (2014, September 2). Notes on the Celebrity Data Theft [Web log post]. Retrieved from https://www.nikcub.com/posts/notes-on-the-celebrity-data-theft/

Davis, D., Monrose, F., & Reiter, M. (2004). On User Choice in Graphical Password Schemes. *This paper is included in the Proceedings of the 13th USENIX Security Symposium.* San Diego, CA.

De Angeli, A. Coventry, L. Johnson, G. & Renaud, K. (2005). Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. International Journal of Human Computer Studies, 63, 128-152.

Ericsson, K. A., & Kintsch, W. (1995). Long-Term Working-Memory. *Psychological Review*, 211-245.

Fletcher, C. R., & Chrysler, S. T. (1990). Surface Forms, Textbases, and Situation Models: Recognition Memory for Three Types of Textual Information. *Discourse Process*, 175-90.

Florencio, D. & Herley, C. (2007). A large-scale study of web password habits. WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.

Gentner, D. (1981). Some interesting differences between verbs and nouns. Cognition and Brain Theory, 4(2), 161-177.

Goldstein, A.G. & Chance, J.E. (1970). Visual recognition memory for complex configurations. Attention, Perception, & Psychophysics, 9(2), 237-241.

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. Journal of Experimental Psychology, 61(1), 23-29.

Hoover, C., Minard, N., Hsu, T., & Werner, S. (2013) *Using an Adaptive Recognition Approach to Measure Information Retention in Story-based Passwords.* Poster session presented at the Northwest Cognition and Memory Conference, Vancouver B.C.

Jacko, J. A. (2012). *The Human-Computer Interaction Handbook.* Boca Raton: CRC Press.

Jain, A.K., Bulle, R. and Pankanti, S. (1999). *Biometrics: Personal Indentification in Networked Society*. Norwell, MA: Kluwer.

Johnson, K., & Werner, S. (2013). Computer Security and Human Memory - Optimizing Password Systems for Users. *In Press*.

Johnson, K. & Werner, S. (2006). Using Composite scene authentication (CSA) as a graphical alternative to alphanumeric password systems. In Proceedings of the 50th annual meeting of the Human Factors and Ergonomics Society. San Francisco, CA. (2006)

Johnson, K. & Werner, S. (2007). Memorability of alphanumeric and Composite Scene Authentication (CSA) passcodes over extended retention intervals. In Proceedings of the 51st annual meeting of the Human Factors and Ergonomics Society. Baltimore, MD. (2007)

Johnson, K. & Werner, S. (2008). Graphical User Authentication – A comparative evaluation of Composite Scene Authentication (CSA) vs. three competing graphical passcode systems (Passfaces, VIP, PassPoints). In Proceedings of the 52nd annual meeting of the Human Factors and Ergonomics Society. Baltimore, MD.

Josang, A. AlFayyadh, B. Grandison, T. AlZomai, M. & McNamara, J. (2007). Security Usability Principles for. Proceedings of the Annual Computer Security Applications Conference. Miami Beach.

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.

Kersten, A. & Earles, J. (2004). Semantic context influences memory for verbs more than memory for nouns. Memory & Cognition, 32(2), 198-211.

Kinnell, A & Dennis, S (2012). The role of stimulus type in list length effects in recognition memory. MEMORY & COGNITION, 40(3), 311-325.

Kintsch, W. (1974). *The Representation of Meaning in Memory.* Hillsdale: Lawrence Erlbaum Associates, Inc.

Kintsch, W. (1998). *Comprehension.* Cambridge: Cambridge University Press.

Kishiyama, MM & Yonelinas, AP (2003). Novelty effects on recollection and familiarity in recognition memory. Memory and Cognition, 31(7), 1045-1051.

Kuo, C., Romanosky, S., & Cranor, L.F. (2006). Human Selection of Mnemonic Phrase-based Passwords. Symposium on Usable Privacy and Secuirty (SOUPS), July 12 – 14, 2006, Pittsburgh, PA, USA.

Kurzban, S. A. (1985). Easily Remembered Passphrases—A Better Approach. ACM SIGSAC Review, 3(2-4), 10-21.

Malmberg, KJ & Nelson, TO (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. MEMORY & COGNITION, 31(1), 35-43.

Maratos, EJ, Allan, K, & Rugg, MD (2000). Recognition memory for emotionally negative and neutral words: an ERP study. NEUROPSYCHOLOGIA, 38(11), 1452-1465.

Medina, J. J. (2008). The Biology of Recognition Memory. Psychiatric Times, June 2008, 13-16.

Nobel, PA & Shiffrin, RM (2001). Retrieval processes in recognition and cued recall. JOURNAL OF EXPERIMENTAL PSYCHOLOGY-LEARNING MEMORY AND COGNITION, 27(2), 384-413.

O'Gorman, L. (2003) Comparing passwords, tokens, and biometrics for user authentication. Proceedings of the IEEE, 91(12), 2019-2020.

Paivio, A. (1963). Learning of adjective-noun paired associates as a function of adjective-noun word order and noun abstractness. Canadian Journal of Psychology/Revue canadienne de psychologie, 17(4), 370-379.

Pavio, A. (1967). Paired-associate learning and free recall of nouns as a function of concreteness, specificity, imagery, and meaningfulness. Psychological Reports, 20(1), 239-245.

Paivio, A. (1986). Mental representations: a dual coding approach. Oxford. England: Oxford University Press.

Passfaces Corporation (2005). *Passfaces™ web access.* Retrieved October, 11th, 2014 from http://www.passfaces.com/products/webaccess.htm

Fisk, Peter (2011). "Ring Worlds". Retrieved September, 2012 from http://www.flashfictiononline.com/

Prabhakar, S., Pankanti, S., and Jain, A.K. (2003). Biometric Recognition: Security and Privacy Concerns. *IEEE Security & Privacy, vol. 1*, no. 2, pp. 33-42.

Radvansky, G. A., & Zacks, R. T. (1991). Mental Models and Fact Retrieval. *Journal of Experimental Psychology Learning Memory and Cognition*, 940-53.

Radvansky, G. A., Spieler, D. H., & Zacks, R. T. (1993). Mental Model Organization. *Journal of Experimental Psychology Learning Memory and Cognition*, 95-114.

Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., & D'Esposito, M. (2004) Dissociable correlates of recollection and familiarity within the medial temporal lobes. Neuropsychologia, 42(1), 2-13.

Renaud, K. & De Angeli, A. (2004). My password is here! An investigation into visuo-spatial authentication mechanisms. Interacting With Computers, 16, 1017-1041.

Rodgers, S. M., Myers, C. W., Ball, J., & Freiman, M. D. (2013). Toward a Situation Model in a Cognitive Architecture. *Computational and Mathematical Orginizational Theory*, 313-345.

Schweitzer, D., Boleng, J., Hughes, C. & Murphy, L., (2009) Visualizing keyboard pattern passwords, Visualization for Cyber Security, 2009. VizSec 2009. 6th International Workshop on , vol., no., pp.69,73, 11-11

Seamon, J. G. (1972). Imagery codes and human information retrieval. Journal of Experimental Psychology, 96(2), 468-470.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior, 6, 156-163.

Standing, L. (1973). Learning 10,000 pictures. Quarterly Journal of Experimental Psychology, 25(2), 207-222.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80(5), 352-373.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse and Comprehension.* New York: New York Academy Press.

Vu, Kim-Phuong L., Proctor, Robert, Bhargav-Spantzel, Abhilasha. Tai, Bik-Lam, Cook Joshua, & Schultz, Eugene (2007) Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies.* 65(8) 744-757.

Wright, N., Patrick, A., & Biddle, R. (2012). Do You See Your Password? Applying Recognition to Textual Passwords. *Symposium on Usable Privacy and Security*, (pp. 1-14). Washington D.C.

Yonelinas, A.P. (2001). Components of episodic memory: the contribution of recollection and familiarity. *Philos. Trans. R. Soc. Lond. B Biol. Sci. 356* (1413), 1363–1374.

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. Journal of Memory and Language, 46, 441-517.

Ziming, Z., Gail-Joon, A., Jeong-Jin, S., & Hongxin, H. (2013). On the Security of Picture Gesture Authentication. *This paper is included in the Proceedings of the 22nd USENIX Security Symposium, pp, 383-398.* Washington, D.C.

Zviran, M. & Haga, W. J. (1993) A Comparison of Password Techniques for Multilevel Authentication Mechanisms. *The Computer Journal*, 36(3) 227-237.

Appendix A

Figure A.1. CSA image version 1, 2, and 3.

Appendix B

B.1 Story Version 1

Elizabeth English had just poured herself a **glass of coffee** and flipped open a book she'd been looking forward to reading when a sudden whooshing sound made her look up in time to witness a **Fairy** materializing in the **library**. For a moment **Elizabeth** considered running out of the **library** and then down the road to the church to fetch **Father Archer** – but she soon dismissed the idea. Her familiarity with otherworldly creatures such as this **Fairy** was poor indeed, but she realized that the **Fairy** might very well take offence at **Father Archer** bursting in, waving a crucifix in its face. Under no circumstances would she have her home turned into a battlefield. Furthermore, **Elizabeth** was not sure that **Father Archer** had any more experience than she did. Was a crucifix the best procedure for dealing with a **Fairy**? **Elizabeth** decided that trying to figure out the answer was taking too much time, so instead she cleared her throat and simply said: "Good afternoon!"

The **Fairy** aimed two pairs of eye-stalks in her general direction, studied her with an inscrutable expression in its hideous face and took two long strides towards her. "Is there anything I can get you?" **Elizabeth** asked putting her **coffee** down with a trembling hand. The **Fairy** took another stride forward, and pointed a claw-like finger at her chest. **Elizabeth** shielded her heart with her hand, but then she realized something and sighed with relief. "Oh, is this what you want?" she said, offering up her **wallet**. The **Fairy** seemed to agree. "I'll take that as a yes." **Elizabeth** held out the **wallet** and the **Fairy** snatched it from her outstretched hand. Then it reached out again.

"You want something more?" But already as she spoke **Elizabeth** realized that she'd misunderstood the gesture. The **Fairy** meant in fact to give her something. "Oh, a barter? My **wallet** for this – **necklace**?" **Elizabeth English** took the pale green object from the **Fairy's** outstretched hand. It felt strange to the touch, as if coated with something that wouldn't allow her skin to come into contact with it. She also noted that it had some small bumps on it, which on closer examination turned out to be 13 tiny **spinners**. A sudden rush of air made her look up, only to discover that the **Fairy** had vanished, as abruptly as it had appeared. She let out another sigh of relief. Admittedly, the **Fairy** had behaved in a quite civilized manner, but she was still happy to see it go. Now she could focus her attention on examining the strange **necklace**.

After twisting and turning it for a while, she decided to simply put it on. It appeared to be too small at first, but to her surprise gently pulling made the necklace widen until she could

slide over her head quite easily. She was equally surprised to discover that her head began vanishing in thin air. She jerked the **necklace** off, and the top of her head reappeared instantly. “Ah, I see.” In exchange for her **wallet** the **Fairy** had given her a **necklace** with the power to make its wearer invisible — or more precisely, to make the wearer invisible from the neck up, a limitation that probably had a somewhat inhibiting effect on the number of possible uses. “Unless — ”She held the **necklace** between her thumbs and index fingers and started pulling it gently. It widened smoothly. Soon she’d made an aperture large enough to pass through. Taking a deep breath **Elizabeth English** stuck her head through the **necklace**.

At once she realized she’d altogether misapprehended the nature of the **necklace**. Before her stretched an endless plain, covered with high, red grass that swayed in a calm breeze. A huge, orange-colored sun balanced on the horizon, and three small, silver-white moons chased each other across a purple sky where alien stars glowed. There was a faint **rotten eggs** smell in the air. She climbed out of it and felt the soft grass under her feet. Straightening her back she clutched the **necklace** firmly to her chest, suddenly afraid that it might just disappear, like something out of a dream. She felt the small **spinners** under her fingers and turned them slowly, thoughtfully as she gazed across the billowing plain, contemplating the implications and the seemingly endless possibilities.

A sudden, irresistible desire for a cup of **coffee** compelled **Elizabeth** to climb into the **necklace** again. Upon sticking her head through the **necklace** again however, she found not her **library** as she expected, but yet another fantastic, alien world. Pulling her head back through, she put her ear close to the **necklace**, turned a **spinner** a notch and heard the faint, ominous click. A cold hand squeezed her heart. She got down on her knees and began drawing numbers on the ground. When she was finished she stood up and stared in dismay at the calculation. 13 spinners and 5 positions for each meant 1,220,703,125 possible combinations. Since she did not have a watch she couldn’t tell the exact time, but she suspected she would be late for work in the morning.

B.2 Story Version 2

Olive O’Neal had just poured herself a **glass of sherry** and flipped open a book she’d been looking forward to reading when a sudden whooshing sound made her look up in time to witness a **Ghost** materializing in the **bedroom**. For a moment **Olive** considered running out of the **bedroom** and then down the road to the church to fetch **Father Archer** – but she soon

dismissed the idea. Her familiarity with otherworldly creatures such as this **Ghost** was poor indeed, but she realized that the **Ghost** might very well take offence at **Father Archer** bursting in, waving a crucifix in its face. Under no circumstances would she have her home turned into a battlefield. Furthermore, **Olive** was not sure that **Father Archer** had any more experience than she did. Was a crucifix the best procedure for dealing with a **Ghost**? **Olive** decided that trying to figure out the answer was taking too much time, so instead she cleared her throat and simply said: "Good afternoon!"

The **Ghost** aimed two pairs of eye-stalks in her general direction, studied her with an inscrutable expression in its hideous face and took two long strides towards her. "Is there anything I can get you?" **Olive** asked putting her **sherry** down with a trembling hand. The **Ghost** took another stride forward, and pointed a claw-like finger at her chest. **Olive** shielded her heart with her hand, but then she realized something and sighed with relief. "Oh, is this what you want?" she said, offering up her **necktie**. The **Ghost** seemed to agree. "I'll take that as a yes." **Olive** held out the **necktie** and the **Ghost** snatched it from her outstretched hand. Then it reached out again.

"You want something more?" But already as she spoke **Olive** realized that she'd misunderstood the gesture. The **Ghost** meant in fact to give her something. "Oh, a barter? My **necktie** for this – **sash**?" **Olive O'Neal** took the pale green object from the **Ghost's** outstretched hand. It felt strange to the touch, as if coated with something that wouldn't allow her skin to come into contact with it. She also noted that it had some small bumps on it, which on closer examination turned out to be 13 tiny **tuners**. A sudden rush of air made her look up, only to discover that the **Ghost** had vanished, as abruptly as it had appeared. She let out another sigh of relief. Admittedly, the **Ghost** had behaved in a quite civilized manner, but she was still happy to see it go. Now she could focus her attention on examining the strange **sash**.

After twisting and turning it for a while, she decided to simply put it on. It appeared to be too small at first, but to her surprise gently pulling made the sash widen until she could slide over her head quite easily. She was equally surprised to discover that her head began vanishing in thin air. She jerked the **sash** off, and the top of her head reappeared instantly. "Ah, I see." In exchange for her **necktie** the **Ghost** had given her a **sash** with the power to make its wearer invisible — or more precisely, to make the wearer invisible from the neck up, a limitation that probably had a somewhat inhibiting effect on the number of possible uses. "Unless — "She held the **sash** between her thumbs and index fingers and started pulling it gently. It widened

smoothly. Soon she'd made an aperture large enough to pass through. Taking a deep breath **Olive O'Neal** stuck her head through the **sash**.

At once she realized she'd altogether misapprehended the nature of the **sash**. Before her stretched an endless plain, covered with high, red grass that swayed in a calm breeze. A huge, orange-colored sun balanced on the horizon, and three small, silver-white moons chased each other across a purple sky where alien stars glowed. There was a faint **dead fish** smell in the air. She climbed out of it and felt the soft grass under her feet. Straightening her back she clutched the **sash** firmly to her chest, suddenly afraid that it might just disappear, like something out of a dream. She felt the small **tuners** under her fingers and turned them slowly, thoughtfully as she gazed across the billowing plain, contemplating the implications and the seemingly endless possibilities.

A sudden, irresistible desire for a cup of **sherry** compelled **Olive** to climb into the **sash** again. Upon sticking her head through the **sash** again however, she found not her **bedroom** as she expected, but yet another fantastic, alien world. Pulling her head back through, she put her ear close to the **sash**, turned a **tuner** a notch and heard the faint, ominous click. A cold hand squeezed her heart. She got down on her knees and began drawing numbers on the ground. When she was finished she stood up and stared in dismay at the calculation. 13 tuners and 5 positions for each meant 1,220,703,125 possible combinations. Since she did not have a watch she couldn't tell the exact time, but she suspected she would be late for work in the morning.

B.3 Story Version 3

Mercedes Meza had just poured herself a **glass of milk** and flipped open a book she'd been looking forward to reading when a sudden whooshing sound made her look up in time to witness a **Martian** materializing in the **living room**. For a moment **Mercedes** considered running out of the **living room** and then down the road to the church to fetch **Father Archer** – but she soon dismissed the idea. Her familiarity with otherworldly creatures such as this **Martian** was poor indeed, but she realized that the **Martian** might very well take offence at **Father Archer** bursting in, waving a crucifix in its face. Under no circumstances would she have her home turned into a battlefield. Furthermore, **Mercedes** was not sure that **Father Archer** had any more experience than she did. Was a crucifix the best procedure for dealing with a **Martian**? **Mercedes** decided that trying to figure out the answer was taking too much time, so instead she cleared her throat and simply said: "Good afternoon!"

The **Martian** aimed two pairs of eye-stalks in her general direction, studied her with an inscrutable expression in its hideous face and took two long strides towards her. “Is there anything I can get you?” **Mercedes** asked putting her **milk** down with a trembling hand. The **Martian** took another stride forward, and pointed a claw-like finger at her chest. **Mercedes** shielded her heart with her hand, but then she realized something and sighed with relief. “Oh, is this what you want?” she said, offering up her **jacket**. The **Martian** seemed to agree. “I’ll take that as a yes.” **Mercedes** held out the **jacket** and the **Martian** snatched it from her outstretched hand. Then it reached out again.

“You want something more?” But already as she spoke **Mercedes** realized that she’d misunderstood the gesture. The **Martian** meant in fact to give her something. “Oh, a barter? My **jacket** for this – **coin**?” **Mercedes Meza** took the pale green object from the **Martian’s** outstretched hand. It felt strange to the touch, as if coated with something that wouldn’t allow her skin to come into contact with it. She also noted that it had some small bumps on it, which on closer examination turned out to be 13 tiny **gears**. A sudden rush of air made her look up, only to discover that the **Martian** had vanished, as abruptly as it had appeared. She let out another sigh of relief. Admittedly, the **Martian** had behaved in a quite civilized manner, but she was still happy to see it go. Now she could focus her attention on examining the strange **coin**.

After twisting and turning it for a while, she decided to simply put it on. It appeared to be too small at first, but to her surprise gently pulling made the coin widen until she could slide over her head quite easily. She was equally surprised to discover that her head began vanishing in thin air. She jerked the **coin** off, and the top of her head reappeared instantly. “Ah, I see.” In exchange for her **jacket** the **Martian** had given her a **coin** with the power to make its wearer invisible — or more precisely, to make the wearer invisible from the neck up, a limitation that probably had a somewhat inhibiting effect on the number of possible uses. “Unless — ”She held the **coin** between her thumbs and index fingers and started pulling it gently. It widened smoothly. Soon she’d made an aperture large enough to pass through. Taking a deep breath **Mercedes Meza** stuck her head through the **coin**.

At once she realized she’d altogether misapprehended the nature of the **coin**. Before her stretched an endless plain, covered with high, red grass that swayed in a calm breeze. A huge, orange-colored sun balanced on the horizon, and three small, silver-white moons chased each other across a purple sky where alien stars glowed. There was a faint **apples** smell in the air. She climbed out of it and felt the soft grass under her feet. Straightening her back she

clutched the **coin** firmly to her chest, suddenly afraid that it might just disappear, like something out of a dream. She felt the small **gears** under her fingers and turned them slowly, thoughtfully as she gazed across the billowing plain, contemplating the implications and the seemingly endless possibilities.

A sudden, irresistible desire for a cup of **milk** compelled **Mercedes** to climb into the **coin** again. Upon sticking her head through the **coin** again however, she found not her **living room** as she expected, but yet another fantastic, alien world. Pulling her head back through, she put her ear close to the **coin**, turned a **gear** a notch and heard the faint, ominous click. A cold hand squeezed her heart. She got down on her knees and began drawing numbers on the ground. When she was finished she stood up and stared in dismay at the calculation. 13 gears and 5 positions for each meant 1,220,703,125 possible combinations. Since she did not have a watch she couldn't tell the exact time, but she suspected she would be late for work in the morning.

Appendix C

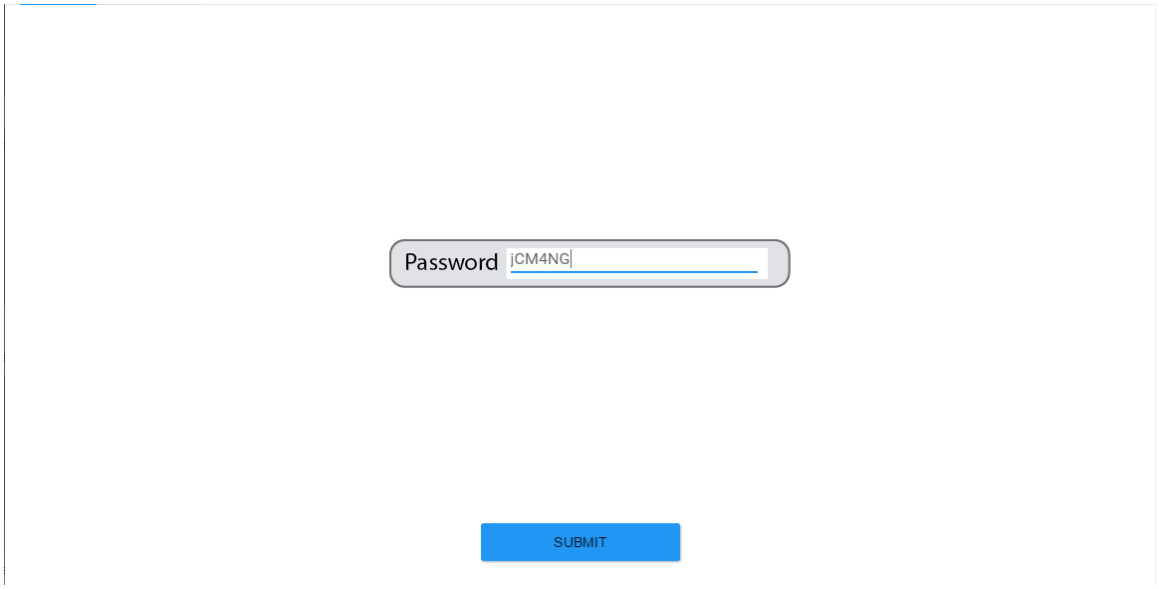


Figure C.1. Alphanumeric authentication screen in Cognilab as seen by the participants.

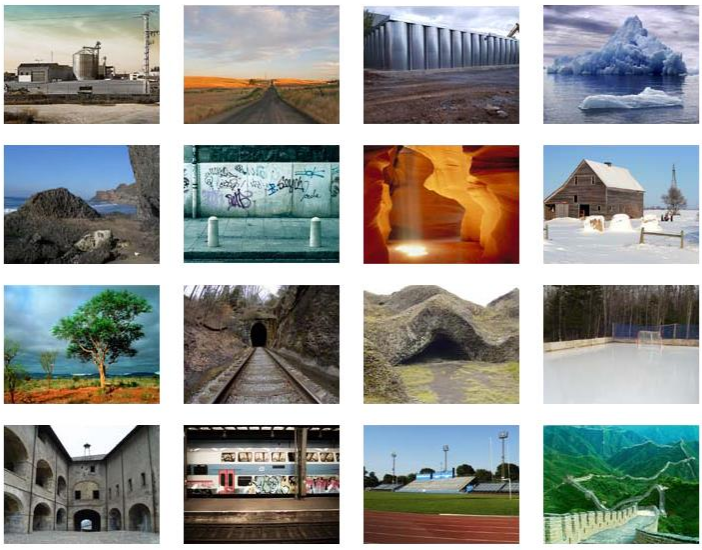


Figure C.2. CSA authenticaction screen 1 in Cognilab as seen by the participants.

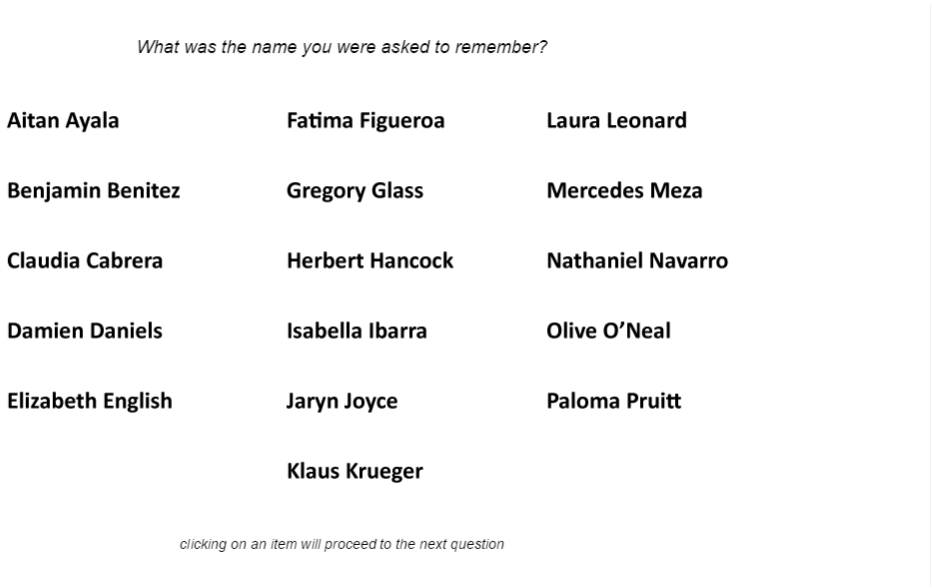


Figure C.3. Narrative password authentication screen 1 in Cognilab as seen by the participants.