

CPasswords: Leveraging Episodic Memory and Human-Centered Design for Better Authentication

Aligning Human and Technical Requirements for Passwords

Prof. L. Jean Camp
School of Informatics and Computing
Indiana University Bloomington

Jacob Abbott
School of Informatics and Computing
Indiana University Bloomington

Abstract

Authentication using passwords requires three cognitively difficult actions. First, a good password requires generation of a high level of entropy. Then the person must reliably recall that highly entropic password. Finally, the person must properly map the password to the context. The common implementation and interaction designs of password studies increase these challenges. We present a system that offers entropy to the user by providing randomly-selected visuals, leverages this source of entropy for password creation, and then further utilizes this visual cue to simplify contextualization. While images have long been used for contextualization, this system is distinct in two ways: the use of images for textual password generation and for textual recall. Our results show a significant increase in entropy and length of passwords created using multiple measures of entropy with no decrease in recall of these more entropic passwords.

1. INTRODUCTION

As long as there are passwords, there will be thefts of passwords including brute force, phishing, and man-in-the-middle attacks. The ease of brute force attacks is grounded in the use of simple, easy to attack passwords. The scope and potential cost of lost passwords is exacerbated by the re-use of passwords, or the use of the same simple algorithms (i.e., domain names with simple substitutions).

There are arguments that the appropriate response to these weaknesses is the removal of passwords all together. The common methods of authentication are something you know (e.g., a passwords), something you have (e.g., a phone), or something you are (biometrics). The concept of social authentication, which builds upon someone you know, is an increasing popular approach.

Yet passwords, “something you know”, dominate authentication. Social authentication, also phrased as “someone you know”, has grown in the last decade with the rise of Web 2.0. Despite the rise of mobile phones as tokens and social authentication, passwords remain. Rather than endeavoring to end the password, we have developed a system to make the password inherently more usable. Such usability is more than interface deep, but requires aligning the core design with human cognitive abilities and heuristics.

The playing field favors attackers in terms of passwords. Entropy is difficult for humans. When looking at a site, individuals are told to think of something random, to not write it down, remember it, and never reuse it. Further, people are not to think of the site name, the word “password”, the keyboard, or the domain name on which the person is gazing. This is the technical equivalent of “don’t think of a pink elephant”. That is, it is unaligned with human behaviors.

In this work we describe a system that has four primary innovations. First, it is aligned with, instead of in opposition to, human cognitive processes. Second, the very process of using photos to offer more randomness in passwords enables the creation of a cue that is linked to a site. We distinguish this from SiteKey below [3]. Third, the system is designed so that the connection between that cue and the password makes it easier for victims to contextualize the password. Specifically, the use of the cue simplifies recall. In addition, successful phishing requires that the target does not closely exam the website. Phishing is made easier by attackers because of habituation.

Here we focus on first two, the creation and recall of passwords. We present experimental evidence that the first two design goals have been met. We begin with a brief overview of the decades of scholarship on passwords. We then detail the system design, followed immediately by the experiment; the fifth section gives results with analysis. We conclude with suggestions for future work.

2. RELATED WORK & MOTIVATION

In the next paragraphs we provide related work on usable passwords. Before providing information on the system itself, we must address the primary objection we have encountered to this system: passwords have been declared dead by more than

one scholar or business. Yet passwords are the most widely used form of authentication [REF], and at least here for the medium term. While some people prefer hardware, such as phones, there are reasons why passwords will remain for some time. Indeed, initial registration of phones requires entering a password at least once. Here we provide four reasons.

First, passwords are low-cost and require no specialized hardware. SecurityKey costs \$15 per person, and are infeasible for companies with large numbers of online customers, low margins, and in some cases high turnover. Secure USBs range from \$6 to \$50 [22].

Second, even with the use of phones, individuals use passwords on desktops and on shared devices. Desktops are over 20% of newly purchased computing devices [25].

Third, people use passwords as de facto access control to share accounts. The most widely documented cases are in the workplace, where security policies prevent individuals from completing their primary task (or are perceived as doing so) [9].

Fourth, passwords are human scale and comprehensible. Individuals feel in control. Extensive studies of risk perception have shown that the perception of control makes risk more acceptable, and further than these perceptions apply to online risk as well as offline risk. Individuals can change their passwords at will, in addition to being able to select their own investment in passwords perceived as being low or high value. The ability to mitigate a risk if one is exposed to the risk, also decreases risk perception, and increases willingness to take a risk.

Multiple studies have looked into the problem of users creating passwords that are easily guessable by other people or are relatively easy to crack [4, 5]. (Cracking refers to the use of automated guessing of common passwords. There are many readably available cracking tools.) Devillers' study [6] tested a large set of passwords to test the state of affairs and suggested that over 90% of those they tested were insecure. Such an alarming amount brings forth questions regarding why users would create passwords with such weaknesses. Tam et al. [7] found that there is a relationship for users between the strength of a password they create and its ease of use.

Usability in authentication design is suggested by Braz and Robert [8] to help decrease the cost for users associated with security systems.

Password strength bars are widely used. Password bars do provide some risk communication and feedback, increasing passwords strength [10]. Risk communication online is a promising area of research [17]. However it is not designed to make the creation of passwords more simple, nor does risk communication simplify recall. Such password bars are not designed to ease creation of passwords, and no study has proven that result.

Automatic password generation is an alternative [11].

Other researchers have advocated graphical passwords [12]. Yet research on graphical passwords has shown that these, while more acceptable to users, generally decrease entropy. Warkentin et al. suggested a password system that used a new input method

for user created passwords that removed the requirement of physically typing in a chosen password [26].

Another alternative is social passwords, where individuals distinguish faces for their social networks from the faces of strangers [13].

We propose that reducing the cognitive load on a user while recalling a password may rest in the type of memory used to store the password. Episodic memory has been shown to be capable of helping individuals' with recall and contextualization. As individuals age, episodic memory does not decline, so that older adults maintain high retention [14, 15]. Visual cues to trigger episodic memory have been found to be more effective than simple text based cues [16, 17] and assist in making technology more accessible [18].

Passmark is the closest technology to CPasswords. There are two difference. CPasswords increases the length and entropy of passwords. Also unlike Passmark the Cognitive Passwords system reduces the requirements on the individual for contextualization. Passmark does not cue the password; thus it requires the user to recall an image unrelated to the authenticating phrase. As opposed to additional unrelated information (answers to questions and an unrelated image), CPasswords' image is integrated into password creation. The image is not functioning solely as a means of authentication; but rather as a way to assist the individual with passphrase recall. Here we illustrate that CPasswords increases recall of longer and more entropic passwords.

3. SYSTEM DESIGN

Cognitive Passwords leverages the use of episodic memory to make construction and recall of passwords easier. In addition it is designed to support creation of passwords with higher entropy, as opposed to simplifying the creation of the same level of passwords.

In order to simplify the creation of passwords the system provides random visual cues. Specifically, the user is asked to create a sentence or string from the photos. This makes the use of the other present cues (domain name, the concept of passwords, and domain name) easier to avoid.

The system generates a random set of five pictures and asks individuals to construct their passwords from those pictures. Humans are a bad source of entropy, and demanding that they behave randomly does not make it possible. With Cognitive Passwords, entropy is provided by the size of the photo library. Specifically, the randomness comes from the photos selected and the variance among individuals selecting descriptions.

Each person can choose one of the menu of images to provide a reminder when they attempt to login. The image shows up after entry of the username, always displayed in the same location relative to the required password entry. This visual reminder will cue and assist in the recall of the correct password. Simon in his canonical work, notes "memory is usually described as 'associative' because of the way in which one thought retrieved from it leads to another. Information is stored in linked list structures [19]." By providing a cue to begin the linked information, CPasswords simplifies recall.

The system is designed so that if an attacker leaves out, misplaces, or selects the wrong image, the victim will have an incorrect cue. This will increase cognitive difficulty of recall, as the person tries to recall a passphrase in the wrong context. This is familiar to anyone who has tried to recognize a colleague who was met in one place, and then encountered in another. The increased difficulty can increase probability that the victim does not simply enter the passphrase via rote behavior. Thus the cue is also designed to make the recollection of the passphrase in the wrong context more difficult.

In the following experiment we provide experimental proof of the increased entropy and ease of recall in context. The experiment was approved by the local IRB.

4. EXPERIMENT DESIGN

We have three claims about the system design: ease of creation, ease of recall, and increased resilience to phishing due to cued recall. Here we describe an experiment to test the first of these. Specifically, we test the two hypotheses.

H₁: Visual cues will lead to the creation of longer passwords with higher entropy.

In order to do this we provide a set of password creation options, breaking individuals into four groups. We compare the entropy of the passwords created in the four scenarios as described below, comparing different types of visual cues with current best practice. Entropy is calculated according to the NIST guidelines [20].

H₂: Visual cues will aid in passphrase recall so that higher entropy passwords are easier to recall.

To test this we used two experiments. In one experiment, we used four control groups. One was provided a common password interface: there is a rule for minimum length and a password compliant with that rule is created. In one case, the rule is enhanced to require the use of at least a single uppercase letter, lowercase letter, number, and a special symbol during password creation. In another both the stronger rule requirements and the images are provided for password creation. In the fourth group, CPasswords is used as designed: stronger rule, images to enable password creation, and a cue for recall.

In order to investigate these hypotheses we deployed an experiment using Amazon's Mechanical Turk. Amazon's Mechanical Turk (MTurk) is an online marketplace where businesses or individuals contract for specific tasks as "Requesters", the individuals complete the "Human Intelligence Tasks" (HITS). These are in contrast to computation tasks, which do not require a human to complete [21].

Participation in our task required MTurk workers have a masters qualification. Amazon rewards master qualifications to workers based on their previous ratings, how long they have had a worker account, the number of hits performed, and completion rates.

We also requested workers have a familiarity with online environments, but not necessarily security conscious. This is likely to describe most MTurk workers.

The initial task was a survey that asked lifestyle questions which were not intended to prime for privacy. Of the initial survey participants, 380 of those chose to create a single password or passphrase. Then the participants were required to return and re-enter the passphrase. Each participant was allowed four attempts to recall and re-enter the password. We distinguish between those who forgot usernames as opposed to passwords.

This request for passwords leveraged the research group tradition of creating desirable hits. The ratings for our requester account are 4.26 out of 5 for communication, 4.68 for pay, 4.88 for being fast, and 4.95 for fairness. In addition to paying quickly and our policy of responding to those queries made during US working hours within an hour or less (ideally immediately). During an experiment we have a designated researcher to monitor the associated email account during working hours from 8am Eastern to 6pm Pacific. We price each hit based on estimated time such that the worker receives the minimum wage. We estimate time by having all members of the research group complete the task on MTurk. (All research group members have a special qualification allowing a limited HIT release.) Participants in the initial task were not required to participate in the follow-up task. Participation in the follow-up task required recollection of a password.

Recall we had 380 original responses for the survey creation component. We filtered responses based on time to completion. In addition, we rejected 30 responses as not legitimate participants due to attempts to submit the completion code of the survey without an actual survey having been completed.

Recall the putative task was a survey unrelated to security, privacy or passwords. The purpose of the unrelated survey was to prevent priming participants to be aware of security as a task. MTurk workers are not passive participants, but have active discussion spaces, for example mturkforum.com. We monitored these discussion spaces to ensure that MTurk participants were not discussing security as a requirement. Thus the MTurk workers were also not providing priming in these communities.

To test recall we have a limited qualification test, which required that they had completed the first task. Of the original respondents to the first task, we had 250 participants complete the follow up task.

4.1. Password Creation

The experiment had participants separated into four different groups for password creation.

Username:

Email:

Choose a password:

Verify password:

Verify password again:

Done

Figure 1. Screenshot for control group

The control group was given feedback as they created their password in the form of a password strength bar. The purpose of this is to compare Cognitive Passwords with best practice. The only requirement on the password was that the password must be a minimum length of eight characters.

Username:

Email:

Choose a password:

(Must have at least one lowercase letter, uppercase letter, number, and symbol.)

Verify password:

Verify password again:

Done






Figure 2. Screenshot for rule group

The second, referred to the Rule group, had the added requirement that all passwords created must include a minimum of one lowercase letter, uppercase letter, digit, and special character. These are common requirements.

Username

Email

Use the pictures to create a password phrase

new pictures

Password Phrase

Verify Password Phrase

Verify Password Phrase Again

Done

The password must have:

- one lowercase letter
- one uppercase letter
- number
- symbol






Figure 3. Screenshot for photo group

The Photo group were given four random pictures and were required to use a specific special character in the password they created while still having all the previously mentioned requirements.

Username

Email

Use the pictures to create a password phrase
 Select one image as a reminder

new pictures

Password Phrase

Verify Password Phrase

Verify Password Phrase Again

Done

The password must have:

- one lowercase letter
- one uppercase letter
- number
- symbol

Figure 4. Screenshot for reminder group

Finally, the Reminder group had the same requirements as the Photo group, except that users had to choose one of the images they were given to use as a reminder.

4.2. Password Recall

To test recall, we waited seven days after the first task had closed. We then created a custom task and invited only previous participants to take part. The follow-up survey required participants to use the credentials they had previously created to login before answering a second small survey. The second survey asked participants to rate the difficulty of recalling their created username and password and to self-report if they wrote down the credential information they had created. Again by monitoring MTurk discussion space we found that participants of the previous experiment did not discuss security. Rather MTurkers identified the HIT as easy and desirable.

5. RESULTS

Analysis of user information differed in each phase of the experiment.

To measure the strength of user's passwords created during Phase 1, we took a look at two factors: length of the password and entropy of the password. The length of the password solely looked at the number of characters used in the created password. We used the user chosen password entropy model calculated per the NIST standard [20]. Data from Phase 2 viewed the number of users who attempted to login using their previously created credentials and logged all attempts. Six of our 250 participants reported using a password generator during Phase 1. We removed those six datapoints from the results below. Recall that the measurement of password entropy was based on [25]. Our goal is to determine if these groups are different relative to each other. Thus we both used a method published in a highly selective venue, we also used a measure consistently across groups.

We provide three presentations of our results in increasing levels of complexity. First we provide simple graphical results for simple visualization. Second we implement null hypothesis significance testing, reporting those results both compared between groups and between the large-scale password database that was released on February 10, 2015 [23]. The use of this data was approved by the IRB as it was pre-existing data, we stripped all user names, and it is stored in the Scholarly Data Archive, a university high-security repository. Finally we provide a Bayesian analysis to confirm that the results are significantly different. We discuss how the selected Bayesian approach enables this assertion in that section, and refer those unfamiliar with this assertion to Kruschke's work [27].

5.1. Graphical Results

5.1.1. Password Length

We observed that 75% of participants in the Control group created passwords that ranged between eight and fourteen characters in length. Fig. 5 shows a density distribution of the control group's responses. It should be noted that of the responses from the control group there were 9 passwords that consisted solely of lowercase alphas, 41 using only uppercase or lowercase alphas combined with at least one digit, only 1 password consisted only of both uppercase and lowercase

alphas, 15 responses used both lowercase and uppercase alphas with a digit, and finally the remaining 32 used a mix of upper/lowercase alphas, digit, and special characters. Less than one third of the responses from the control group would meet the minimal requirements imposed on our remaining three groups.

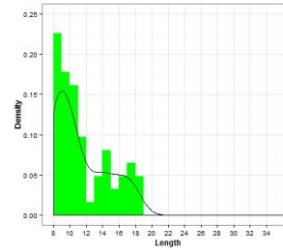


Figure 5. Control length

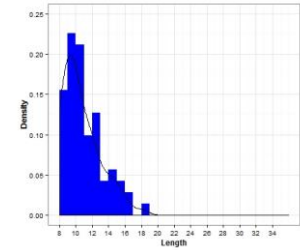


Figure 6. Rule length

The Rule group had 75% between 8 and 12 characters in length. Fig. 6 shows that after the median response of ten characters in length, passwords of longer lengths occurred more in line with a normal distribution than later responses in the Control group.

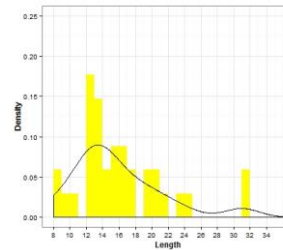


Figure 7. Photo length

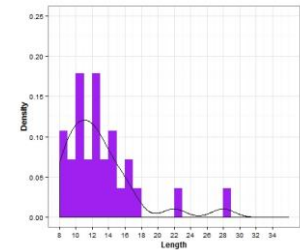


Figure 8. Reminder length

The Photo group had the widest range of responses with 75% creating passwords between 8 and 18 characters in length, but had the longest password at 31 characters. Fig. 7 shows the distribution of the created passwords in this group. Even removing the farthest outlier at 65 characters, the largest password still reached 31 characters in length.

The Reminder group had the second widest range of responses with 75% of users responding with passwords between 8 and 14 characters in length. The longest response in the Reminder group was 28 characters long. See Fig. 8 for the distribution of the Reminder group's passwords.

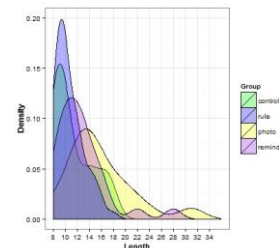


Figure 9. Length distribution

Despite the varying rules and cues, the density of responses across all groups centered closer towards the established minimum length, but extended further to the right with the

Photo and Reminder groups than the Control or Rule groups, as shown in Fig. 9.

5.1.2. Password Entropy

The length of a created password is only one aspect of its complexity, the entropy created by a participant is also a key factor in measuring the strength and resiliency of a password. In Fig. 10 the results of calculating entropy for the Control group shows that 50% of the responses created only 24.75 bits of entropy. This is most notable due to the results of the other three groups who all had a minimum of 24 bits of entropy for their shortest passwords. Even taking the longest password from the Control group, it only created 39 bits of entropy, which is tied with the Rule group as the lowest of all four groups when comparing the maximum values they contain.

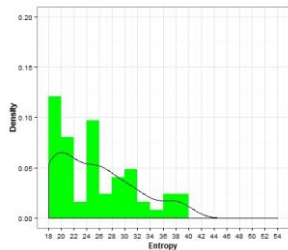


Figure 10. Control entropy

The Rule group had a higher value than the majority of those in the Control group. Three quarters of the passwords created by participants had more than 25 bits of entropy, which outdid the Control group’s median value of 24.75 bits of entropy. Fig. 11 shows a clear shift to the right of the Rule group’s password distribution as compared to the Control group. Similar to the pattern established by looking at the length of the created passwords, the Rule group had second lowest results of the four groups for bits of entropy.

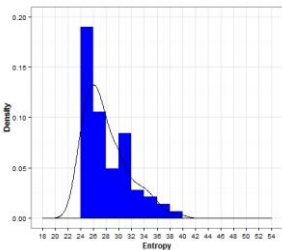


Figure 11. Rule entropy

The Photo group’s results surpassed those of the Control and Rule groups. The median amount of bits created by the Photo group reached 33.75 bits of entropy, higher than the 75% of responses in both the Control and Rule group, which were calculated at most 30 bits of entropy. Fig. 12 shows how much farther right the density curve stretches compared with the other response groups.

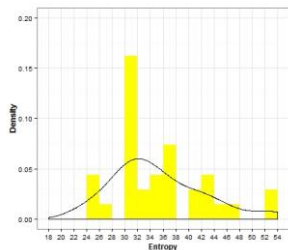


Figure 12. Photo entropy

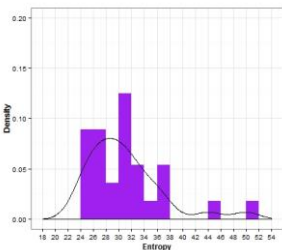


Figure 13. Reminder entropy

Resulting in the third highest set of entropy, the Reminder group fell only behind the Photo group. The median of the Reminder group was 30 bits of entropy as opposed to the Rule group’s median of 33.75 bits or the Control group’s 24.75 bits. Fig. 13 shows how responses stretched farther from the basic minimum area than all but the Photo group.

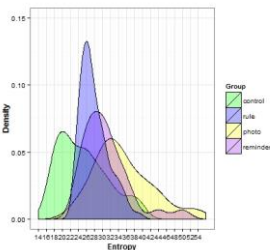


Figure 14. Entropy distribution

Fig. 14 shows the distributions of all four groups overlapped to give an improved visual comparison between groups. Despite the closely overlapped density of password length across groups, the distribution of entropy between groups is more pronounced. Most noticeably is how left skewed the Control group is when compared to the remaining three groups.

Participation in the follow-up task to test subject recall was met with slightly over a 70% participation rate. Each participant had a time delay between the completion of their first task and the beginning of the second task that ranged from one to two weeks in length. The time delay varied by when the tasks were released to workers and when workers individually took and completed the assignment.

5.1.3. Password Recall

Table 1 shows a summary of successful login rates based on each group. The Control group had the highest success rating with 50% successfully logging in with the credentials they had previously created. The Reminder group came in third with a success rating of 43%, just behind the Rule group’s recall rating of 45%. While the Reminder group’s recall rate is lower than those of the Control and Rule groups, the variation between the entropy values of these groups appear significant.

Table 1. Login success rate

Group	Control	Rule	Photo	Reminder
Success Rate	50%	45%	26.5%	43%

Requiring different types of characters to be used and giving visual cues appeared to significantly raise the entropy values of passwords that were created in the visualizations of the respective groups’ data.

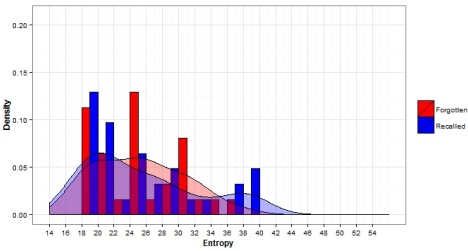


Figure 14. Control recall

Fig. 14 shows the distributions of the Control group by whether participants successfully recalled their passwords or had forgotten between Phase 1 and 2 of the study.

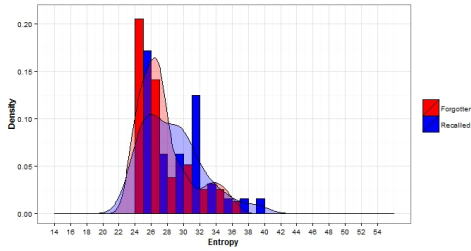


Figure 15. Rule recall

The Rule group's recall distributions, as shown in Fig. 15, are skewed farther to the right than those of the Control group, which is to be expected given that the other groups had requirements that created a higher minimum entropy value for the generated passwords. The Rule group also had the highest density spikes of any group, both located at the Rule group's minimum entropy value.

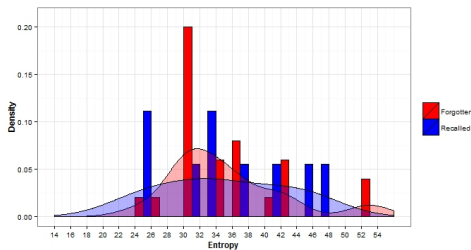


Figure 16. Photo recall

The recall distributions of the Photo group, represented in Fig. 16, have a wider successful recall area with a lower density than any other group. The Photo group had the lowest successful recall rate of any group at only 26.5%.

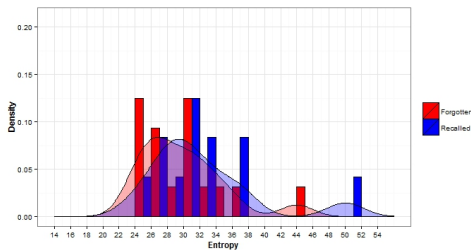


Figure 17. Reminder recall

Fig. 17 shows the recall distributions of the Reminder group, which had higher successful recall than the Photo group, while the likelihood of forgetting the created passwords spiked to lower degrees than in the Rule and Photo group.

The Control group's recall rates reflect the standard expected relationship between entropy and recall, with the forgotten curve being skewed farther to the right over higher values of entropy than the successfully recalled curve.

5.2. Null Hypothesis Testing

Visualizations of the data suggest that differences between the four groups might be significant, so to further evaluate the significance of the data we implement multiple significance tests to test the null hypothesis of there being no significant difference between groups.

5.2.1. Password Length

To compare the length of passwords between groups, we used restricted maximum likelihood estimation of linear mixed effect model to test for significant differences amongst the groups. Table 2 contains the results of our testing and gives the used formula and t-values to show the differences between groups.

Table 2. Group's effect on length

Linear mixed model fit by REML ['lmerMod']

Formula: length ~ group + (1 | counter)

Data: remember

REML criterion at convergence: 1049.4

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.1327	-0.6707	-0.2107	0.4360	4.1470

Random effects:

Groups	Name	Variance	Std.Dev.
counter	(Intercept)	0.5327	0.7298

Residual	Variance	Std.Dev.
	12.6441	3.5559

Number of obs: 195, groups: counter, 71

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	11.2594	0.4610	24.426
grouprule	-0.5974	0.6189	-0.965
groupphoto	4.4991	0.7631	5.896
groupreminder	1.4683	0.8155	1.801

Correlation of Fixed Effects:

	(Intr)	grouprule	groupphoto	groupreminder
grouprule	-0.718			
groupphoto	-0.580	0.432		
groupreminder	-0.543	0.404	0.338	

The results of testing differences of password length between groups showed significant differences between the Control group and both the Photo and Reminder group. The Photo group had a t-value that equated to a significant p-value less than 0.01, while the Reminder group's significant p-value was higher but still less than 0.05.

The Rule group was the only group that did not show a significant difference between the length of participant created passwords and those Control group, with a p-value that was higher than 0.05.

5.2.2. Password Entropy

To evaluate if our experimental control group was representative of the large scale password database that was released. We removed all the passwords contained in the database that contained less than eight characters and took a random sample from the five million remaining passwords. Comparing the entropy of passwords from the control group and the representative sample of the large set resulted in a p-value of 0.19, so we would accept the null hypothesis that our control group is not significantly different from the random sample from the large password set.

Table 3 shows the information from using a restricted maximum likelihood estimation of linear mixed effect model to test for significant differences in the entropy of passwords created between groups.

Table 3. Group's effect on entropy

Linear mixed model fit by REML ['lmerMod']				
Formula: entropy ~ group + (1 counter)				
Data: remember				
REML criterion at convergence: 1211.8				
Scaled residuals:				
Min	1Q	Median	3Q	Max
-2.0100	-0.7057	-0.1732	0.4360	3.4148
Random effects:				
Groups	Name	Variance	Std.Dev.	
counter	(Intercept)	0.3604	0.6003	
Residual		30.4244	5.5158	
Number of obs: 195, groups: counter, 71				
Fixed effects:				
	Estimate	Std. Error	t value	
(Intercept)	25.3070	0.7046	35.91	
grouprule	2.6859	0.9591	2.80	
groupphoto	9.8829	1.1791	8.38	
groupreminder	5.6076	1.2586	4.46	
Correlation of Fixed Effects:				
	(Intr)	grouprl	grpht	
grouprule	-0.727			
groupphoto	-0.591	0.434		
groupremndr	-0.553	0.406	0.334	

The results of testing differences of password entropy between groups showed significant differences between the Control group and both all of the experimental groups. The Photo and Reminder groups had t-values that equated to significant p-values that were less than 0.001, while the Rule group's p-value was significant to a lesser extent at less than 0.01.

5.2.3. Password Recall

To look at whether successful recall was significantly different between groups we used a generalized linear mixed model with maximum likelihood estimation and used the recall information as a binomial factor where participants had either successfully recalled or forgotten their password.

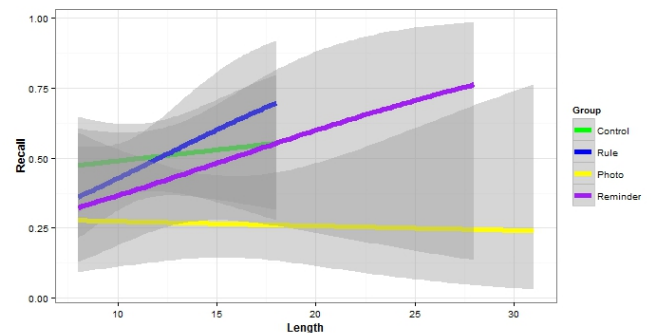
Table 4 shows the information from using the generalized linear mixed model and shows that factors of password length and which group a participant is in are very significant. Each group and factor had a p-value less than 0.001.

To better demonstrate the significance of the results of Table 4, Fig. 18 shows the projected recall by group and length. Each groups' regression line is presented with the probability of successful recall given the length of a password with a 95% confidence interval displayed. For each group the confidence interval is tighter for lower lengths of passwords, where more data is located, while higher lengths on the graph have confidence intervals that widen, projecting less certainty of successful recall.

Table 4. Recall effected by group and length

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: binomial (logit)
Formula: remember ~ entropy + group + (1 | counter)
Data: remember
AIC BIC logLik deviance df.resid
264.9 284.5 -126.5 252.9 189
Scaled residuals:
Min 1Q Median 3Q Max
-1.3884 -0.6911 -0.4159 0.7944 1.6511
Random effects:
Groups Name Variance Std.Dev.
counter (Intercept) 1.27 1.127
Number of obs: 195, groups: counter, 71
Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.695855 0.002366 -294.1 <2e-16 ***
entropy 0.028483 0.002275 12.5 <2e-16 ***
grouprule -0.363171 0.002366 -153.5 <2e-16 ***
groupphoto -1.386081 0.002367 -585.7 <2e-16 ***
groupreminder -0.096564 0.002366 -40.8 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Correlation of Fixed Effects:
(Intr) entpry groprl grppht
entropy -0.003
grouprule 0.000 -0.001
groupphoto 0.000 0.000 0.000
groupremndr 0.000 -0.001 0.000 0.000

**Figure 18. Projected recall by group and length**

Worth noting is the condensed range of the Control and Rule groups' lengths compared to those of the Photo and Reminder groups.

Table 5 shows the information from using the generalized linear mixed model to test for significant differences between recall by a password's entropy and the experimental group the participant was in. Similar to the results testing length, each group and factor had a p-value less than 0.001 and showed that the entropy and groups had a significant impact on the likelihood of recall.

Table 5. Recall effected by group and entropy

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial (logit)

Formula: remember ~ entropy + group + (1 | counter)

Data: remember

AIC	BIC	logLik	deviance	df.resid
264.9	284.5	-126.5	252.9	189

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.3884	-0.6911	-0.4159	0.7944	1.6511

Random effects:

Groups	Name	Variance	Std.Dev.
counter	(Intercept)	1.27	1.127

Number of obs: 195, groups: counter, 71

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.695855	0.002366	-294.1	<2e-16 ***
entropy	0.028483	0.002275	12.5	<2e-16 ***
grouprule	-0.363171	0.002366	-153.5	<2e-16 ***
groupphoto	-1.386081	0.002367	-585.7	<2e-16 ***
groupreminder	-0.096564	0.002366	-40.8	<2e-16 ***

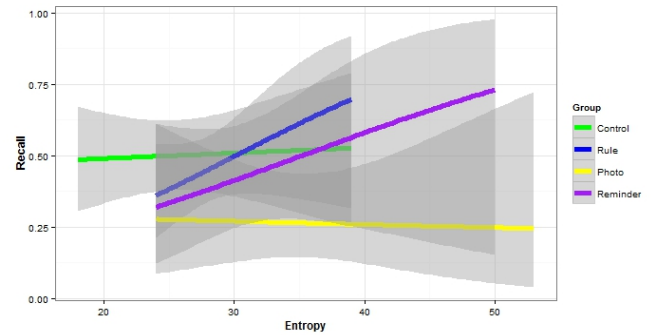
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	entropy	groprl	grpght
entropy	-0.003			
grouprule	0.000	-0.001		
groupphoto	0.000	0.000	0.000	
groupremndr	0.000	-0.001	0.000	0.000

Fig. 19 shows the projected recall by group and entropy to better demonstrate the significance of the results in Table 5. Each groups' regression line is presented with the probability of successful recall given the length of a password with a 95% confidence interval displayed. For each group the confidence interval is tighter for lower entropy values of passwords, where more data is located, while higher values of entropy on the graph have confidence intervals that widen, projecting less certainty of successful recall.

Amongst the four groups, the Rule group actually contained the narrowest range of entropy values of any group. The Photo group had the widest range of entropy values, while also maintaining the lowest projected recall rate. The Control group's entropy values contain a sizable portion of range that is lower than the minimum of the other three groups tested. The projected recall of the Reminder group was higher than the Photo group and extends to higher entropy values than the Control and Rule groups.

**Figure 19. Projected recall of groups by entropy**

From the significant values across groups for factors of length, entropy, and recall, we would reject the null hypotheses of visual cues not leading to the creation of longer passwords with higher entropy and that visual cues would not assist in recalling passwords with higher entropy.

5.3. Bayesian Analysis

The use of NHST to analyze our data suggested many points of significance, but with growing concern over the use of NHST as adequate for the rejection of the null hypothesis and its effectiveness increasingly being contested [24], we therefore use an additional Bayesian approach to analyze our data.

5.3.1. Password Length

To analyze the length of passwords between groups we used a two factor hierarchical Bayesian model and ran twenty-five thousand iterations of Markov Chain Monte Carlo (MCMC) chains, each with fifty thousand steps to create projected posterior distributions of password lengths for each group.

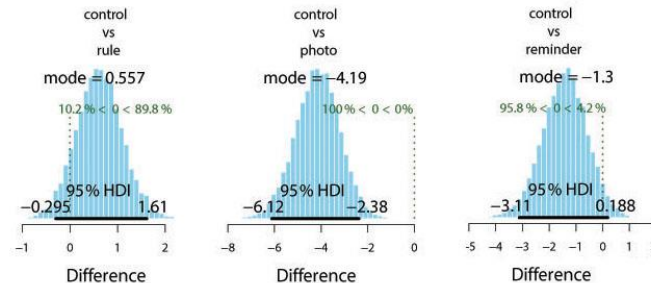
**Figure 20. Length by group comparison**

Fig. 20 presents the comparison of password length distributions between each experimental group and the Control group. The projected comparison of the Control and Rule groups' lengths reflect the results obtained in the previous section as the range of distribution would suggest that password of the Rule group would be the same length or a character shorter.

The Photo group's results show a much more significant difference in distribution from the Control group's length. The comparison reflects the sizable difference between the means of the two groups. The Reminder group's passwords had a mildly significant difference from the Control groups with a range of outcomes from matching the length of a Control password to having an additional three characters.

5.3.2. Password Entropy

To analyze the entropy of passwords between groups we used the same two factor hierarchical Bayesian model and ran twenty-five thousand iterations of Markov Chain Monte Carlo (MCMC) chains, each with fifty thousand steps to create projected posterior distributions of password entropy values for each.

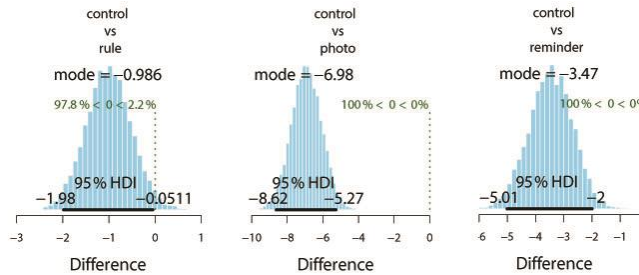


Figure 21. Entropy comparison

The comparisons of password entropy distributions between the Control group and each experimental group are shown in Fig. 21. The projected difference of the Control and Rule groups' entropy values is significant as the distribution's 95% Highest Density Interval (HDI) excludes zero.

The comparison of the Control and Photo groups, as well as the Control and Reminder group comparison, showed significant differences between the entropy values of the groups compared to the Control group.

6. FUTURE WORK

There are three major efforts for future work. The first is to complete the analysis of created and recalled passwords with an improved measure of entropy. Specifically, passwords will be measured by how many guesses are required by simple brute force attacks, rainbow tables, and grammar-aware password crackers. This removes variables such as the processing power of the machine and the quality of the basic brute force attack, providing a better measure. This will precede more complete analysis of the relationships between the experimental groups. The second major effort is an experiment to determine if identical phishing attacks are more or less effective with Cognitive Passwords. Third, and requiring the results of the work above, is an economic model of the costs and benefits of phishing based on these experiments.

7. REFERENCES

- [1] Rachna Dhamija, J. D. Tygar, and Marti Hearst. 2006. Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '06), Rebecca Grinter, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries, and Gary Olson (Eds.). ACM, New York, NY, USA, 581-590.
- [2] Mark Blythe, Helen Petrie, and John A. Clark. 2011. F for fake: four studies on how we fall for phish. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 3469-3478.
- [3] Markus Jakobsson and Steven Myers. 2007. Delayed password disclosure. *SIGACT News* 38, 3 (September 2007), 56-75.
- [4] Dell'Amico, M. (2010). Password strength: An empirical analysis. ... , 2010 *Proceedings IEEE*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5461951
- [5] Florêncio, D. (2007). A Large-Scale Study of Web Password Habits, 657-665.
- [6] Devillers, M. (2010). Analyzing password strength. *Radboud University Nijmegen, Tech. Rep.*(2). Retrieved from http://www.cs.ru.nl/bachelorscripts/2010/Martin_Devillers___0437999___Analyzing_password_strength.pdf
- [7] Tam, L., Glassman, M., & Vandenwauver, M. (2010). The psychology of password management: a tradeoff between security and convenience. *Behaviour & Information Technology*, 29(3), 233-244. doi:10.1080/01449290903121386
- [8] Braz, Christina and Robert, Jean-Marc. Security and usability: the case of the user authentication methods. s.l. : ACM, 2006. pp. 199-203.
- [9] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* 42, 12 (December 1999), 40-46.
- [10] Alain Forget, Sonia Chiasson, P. C. van Oorschot, and Robert Biddle. 2008. Improving text passwords through persuasion. In *Proceedings of the 4th symposium on Usable privacy and security* (SOUPS '08). ACM, New York, NY, USA, 1-12.
- [11] Jeff Yan, Alan Blackwell, Ross Anderson, and Alasdair Grant. 2004. Password Memorability and Security: Empirical Results. *IEEE Security and Privacy* 2, 5 (September 2004), 25-31.
- [12] Robert Biddle, Sonia Chiasson, and P.C. Van Oorschot. 2012. Graphical passwords: Learning from the first twelve years. *ACM Comput. Surv.* 44, 4, Article 19 (September 2012),
- [13] Mohamed Eljelawi, Ali, & Norafida Bt.Ithnin. "Graphical Password: Usable Graphical Password Prototype." *Journal of International Commercial Law and Technology* [Online], 4.4 (2009): 299 - 310.
- [14] Anderson, N. and Craik, F., "Memory in the aging brain", The Oxford handbook of memory, pp. 411-425, 2000.
- [15] Aslan, A., Bauml, K., and Pastotter, B., "No inhibitory deficit in older adults' episodic memory", *Psychological Science*, Vol. 18(1), pp. 72, 2007.
- [16] Herron, C., York, H., Corrie, C., and Cole, S., "A comparison study of the effects of a story-based video instructional package versus a text-based instructional package in the intermediate level-foreign language classroom.", *CALICO Journal*, Vol. 23(2), pp. 281, 2006.
- [17] Vaibhav Garg, L Jean Camp and Kay Connelly, "Risk Communication Design: Video vs. Text", *PETS* (Vigo, Spain) 11-13 July 2012.
- [18] Z. Zimmerman & L Jean Camp, "Elder-friendly Design's Effects on Acceptance of Novel Technologies", *Elderly Interaction Design CHI: CHI 2010 Workshop*, (Atlanta GA.) 4 April 2010.
- [19] Herbert A. Simon. 1996. *The Sciences of the Artificial* (3rd Ed.). MIT Press, Cambridge, MA, USA.
- [20] Burr, W. E., Dodson, D. F., Newton, E. M., Perlner, R. A., Polk, W. T., Gupta, S., and Nabbus, E. A., "Sp 800-63-2. Electronic Authentication Guideline," National Institute of Standards & Technology, Tech. Rep., 2013. <http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-2.pdf>
- [21] <https://www.mturk.com/mturk/welcome>, October 29, 2014.
- [22] Amazon, key word search security key usb, 22 April 2015, http://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Daps&field-keywords=Security+Key+usb
- [23] Burnett, M. 10 million passwords. 10 February 2015, https://xato.net/passwords/ten-million-passwords/?utm_content=buffer545b2#.VNonm75sSS0
- [24] Trafimow, D. and Marks, M. Editorial, Basic and Applied Social Psychology, Vol. 37, Iss. 1, 2015.
- [25] Columbus, Louis. "IDC: 87% Of Connected Devices Sales By 2017 Will Be Tablets And Smartphones." *Forbes*. Forbes Magazine, 12 Sept. 2013. Web. 14 May 2015. <http://www.forbes.com/sites/louiscolumbus/2013/09/12/idc-87-of-connected-devices-by-2017-will-be-tablets-and-smartphones/>
- [26] Warkentin, Merrill, Kimberly Davis and Ernst Bekkering. "Introducing the Check-Off Password System (COPS): An Advancement in User Authentication Methods and Information Security." *JOEUC* 16.3 (2004): 41-58. Web. 13 Jun. 2015. doi:10.4018/joeuc.2004070103
- [27] Kruschke, John K. "Introduction: Credibility, Models, and Parameters." *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd ed. Burlington, MA: Academic/Elsevier, 2014. 15-30. Print.