Correct horse battery staple: Exploring the usability of system-assigned passphrases

Richard Shay, Patrick Gage Kelley, Saranga Komanduri, Michelle L. Mazurek,
Blase Ur, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor
Carnegie Mellon University Pittsburgh, PA
{rshay,pgage,sarangak,mmazurek,bur,tvidas,lbauer,nicolasc,lorrie}@cmu.edu

ABSTRACT

Users tend to create passwords that are easy to guess, while system-assigned passwords tend to be hard to remember. Passphrases, space-delimited sets of natural language words, have been suggested as both secure and usable for decades. In a 1,476-participant online study, we explored the usability of 3- and 4-word system-assigned passphrases in comparison to system-assigned passwords composed of 5 to 6 random characters, and 8-character system-assigned pronounceable passwords. Contrary to expectations, system-assigned passphrases performed similarly to system-assigned passwords of similar entropy across the usability metrics we examined. Passphrases and passwords were forgotten at similar rates, led to similar levels of user difficulty and annoyance, and were both written down by a majority of participants. However, passphrases took significantly longer for participants to enter, and appear to require error-correction to counteract entry mistakes. Passphrase usability did not seem to increase when we shrunk the dictionary from which words were chosen, reduced the number of words in a passphrase, or allowed users to change the order of words.

Categories and Subject Descriptors

D.4.6 [**Security and Protection**]: Authentication; H.1.2 [**User/ Machine Systems**]: Human factors

Keywords

Passphrases, System-assigned passwords, Usability, Password composition policies

1. INTRODUCTION

Passwords are the most common form of authentication, used in both corporate and personal settings. Despite their importance, however, the best approach to using passwords remains an open question. Allowing users free rein to create their own passwords often leads to weak, easily guessed passwords [5, 43, 60], resulting in security breaches and loss of privacy for victims [12].

Many organizations attempt to address this problem using password-composition policies, which limit the password-creation space in an effort to prevent users from choosing passwords that are too easily guessed [10]. Unfortunately, strict password-composition policies sometimes lead to user frustration without substantial security benefit [1, 23]. Also, even under a strict policy, users may fulfill policy requirements in predictable ways [54], such as basing their passwords on older passwords, names, or words [51, 60], or reusing passwords across domains [17].

One approach to making passwords more secure is to remove user choice and have the authentication system generate passwords randomly. Such *system-assigned* passwords can be guaranteed to be sufficiently difficult to guess, although they have been perceived as difficult to remember and type [37].

A *passphrase* is a password composed of a sequence of words. Passphrases are typically much longer than ordinary passwords, and proponents argue that they are more secure and easier to remember. One NIST publication states that "any long password that can be remembered must necessarily be a 'pass-phrase' composed of dictionary words" [10]. The use of passphrases has recently garnered appreciable attention [40, 49], and some institutions have adopted passphrases as a password policy (e.g., [58]). Despite this recent interest in passphrases, however, there is little empirical evidence to support claims of superior usability over passwords.

This paper describes the results of a 1,476-participant study on the usability of system-assigned passphrases and system-assigned passwords. The passphrases we study are sequences of three or four English words drawn at random from a set dictionary and separated with spaces; our passwords are also system-assigned and are five to eight characters in length. We focus on system-assigned, rather than user-selected, passphrases and passwords because this allows us to control for guessability while evaluating usability.

We assigned each participant to one of 11 experimental conditions (three password conditions and eight passphrase conditions). We measured how quickly and accurately participants could enter their password or passphrase both shortly after assignment and several days later. We also asked our participants to complete two brief surveys about their behavior and sentiment.

Our findings suggest that system-assigned passphrases are far from a panacea for user authentication. Rather than committing them to memory, users tend to write down or otherwise store both passwords and passphrases when they are system assigned. When compared to our password conditions, no passphrase condition significantly outperformed passwords in any of our usability metrics, indicating that the system-assigned passphrase types we tested fail to offer substantial usability benefits over system-assigned passwords of equivalent strength. We even find that system-assigned passphrases might actually be less usable than system-assigned passwords. For instance, users were able to enter their passwords more quickly and with fewer errors than passphrases of similar strength.

Symposium on Usable Privacy and Security (SOUPS) 2012, July 11-13, 2012, Washington, DC, USA.

While our results in general do not strongly favor system-assigned passwords over system-assigned passphrases or vice versa, we identify several areas for further investigation. For example, larger dictionary sizes do not appear to have a substantial impact on usability for passphrases. This could be leveraged to make stronger passphrases without much usability cost. We also find that lowercase, pronounceable passwords are an unexpectedly promising strategy for generating system-assigned passwords.

Researchers have proposed error-corrected passphrase systems [3, 25,39]. Our results suggest that sophisticated error correction, such as mapping the word a user enters to the closest word in the passphrase dictionary, is necessary to make passphrases comparably usable to passwords. Without error correction, many passphrase conditions perform significantly worse than our password conditions.

We next discuss the background and related work for our study in Section 2, and turn to our methodology in Section 3. We present results on usability, accuracy, and sentiment in Section 4, and describe our error analysis in Section 5. We consider ecological validity and summarize our findings in Section 6.

2. BACKGROUND AND RELATED WORK

Many years of research have shown that users have difficulty picking strong passwords. Despite a litany of proposed "password replacements" over the past decades, no system has proven superior to text passwords when evaluated according to a broad set of criteria [7, 20]. However, the fact remains that users tend to choose predictable passwords [6, 15] and reuse the same passwords across multiple accounts [17]. Despite warnings to the contrary, many users also write their passwords down [23, 45, 62]. Some researchers argue this practice is not inherently bad [47].

While many user-selected passwords are easily guessed, system-assigned passwords can have much stronger security guarantees; however, they may be difficult for users to remember without writing them down. The literature contains many proposals for making system-assigned passwords easier for users to recall. In this paper, we focus on one such technique, passphrases. We consider a ***passphrase*** to be a set, sometimes ordered, of natural-language words separated by spaces. Passphrases thus tend to be longer than typical passwords. Hereafter, when we use the term ***password***, we refer to a string of approximately 5–16 characters, usually without spaces, that may not have natural-language meaning. We use ***secret*** to refer to the general class of password-like strings containing both passwords and passphrases.

In this section, we first provide background on user-selected passwords, focusing on techniques to make them stronger. We then present related work on techniques for crafting usable, system-assigned secrets. We then discuss user-generated passphrases as encountered "in the wild" and in literature, and, finally, prior studies comparing variations of passwords and passphrases.

2.1 Passwords

Many techniques have been suggested to help users create better passwords. One technique requires that passwords comply with a password-composition policy, such as forcing the inclusion of digits in passwords. However, passwords that comply with such policies often remain vulnerable because policy requirements are sometimes fullfilled in predictable ways [51, 57]. Furthermore, inflexible policies can overly burden users, leading to frustration [23].

Another approach is to allow users to create their own passwords, but proactively check their strength. Prior work has suggested rejecting unsuitable passwords using predefined lists of weak [5] or popular passwords [46]. Other work has suggested informing users of the strength of their proposed password [11, 52, 56].

2.2 System-Assigned Secrets

System-assigned passwords and passphrases have also been studied, but to a lesser degree. Diceware uses manual die rolls to select words for a passphrase from a specialized 7,776-word dictionary [44]. In contrast, early work by Kurzban proposes a passphrase system using only a 100-word dictionary [34]. In our study, we test and compare dictionaries from 181 to 1,024 words in size.

Attempts to create memorable system-assigned passwords include generating passwords that are "pronounceable" by speakers of a natural language. Such passwords generally consist of a series of concatenated syllables from a natural language. Early work by Gasser on generating pronounceable passwords [19] has been adopted in modified form as a standard of the U.S. National Institute of Standards and Technology (NIST) [41]. However, this system uses the frequencies with which syllables appear in English as part of its generation process, which has been shown to increase greatly the guessability of generated passwords [36]. More recent work has proposed other schemes for randomly generating pronounceable text [13], though the usability of these schemes has not been analyzed comprehensively. Jeyaraman and Topkara suggest randomly generating a lower-case password and then automatically creating a mnemonic for that random password in order to make system-assigned passwords more memorable [26].

Other systems use partially system-assigned passphrases and passwords. For instance, Forget et al. insert randomness into user-chosen passwords to increase strength [18]. Lee and Ewe use a variant of this technique to strenghthen user-generated passphrasses by adding random semantic noise [35].

2.3 Passphrases

For three decades, academic literature has considered passphrases as a potentially more memorable and secure alternative to short passwords [10, 27, 42], yet their usability vis-à-vis passwords has not been well studied. Debate about passphrases was recently reignited by the online comic xkcd (Appendix A), which suggested passphrases as an alternative to complex password policies [40]. This comic has been widely reprinted, including in advice to help users create strong and memorable passwords [48].

Despite a lack of empirical evidence, passphrases are recommended by some system administrators. For more than five years, Indiana University has required all new users to create a passphrase, which they define as containing at least 15 characters split among at least four words, delimited by a space or non-underscore symbol.[1] In a recent blog post, the university pointed to the xkcd comic as evidence in favor of this policy [58]. Creighton University [49] and UC Santa Cruz[2] are among other universities that have cited the xkcd comic while suggesting passphrases. Other organizations, such as Clemson University,[3] have suggested that users create passphrases by envisioning sentence-like passphrases. In one of our experimental conditions, we test system-generated passwords constructed grammatically to resemble sentences.

Yan et al. studied mnemonic passwords, in which users select a phrase but type in only the first character of each word [59]. Universities including Carnegie Mellon[4] currently recommend this mnemonic approach on password advice pages. However, Kuo et al. built a dictionary from Internet sources to crack user-generated

[1] `http://kb.iu.edu/data/acpu.html` (visited 5/2012)

[2] `http://its.ucsc.edu/security/training/password.html` (visited 5/2012)

[3] `http://www.clemson.edu/ccit/about/policies/strong_passwords.html` (visited 5/2012)

[4] `http://www.cs.cmu.edu/~help/security/choosing_passwords.html` (visited 5/2012)

mnemonic passwords, suggesting this technique does not prevent users from choosing weak secrets [33]. Bonneau and Shutova examined 100,000 user-selected Amazon.com "payphrases," or globally unique multi-word secrets. They find that lists of popular movies and books, as well as digrams popular in natural language, are effective in guessing these payphrases [8]. Both these works find that user-chosen passphrases may be more easily guessed than expected, which supports the investigation of system-assigned passphrases as an alternative.

2.3.1 Comparative Evaluation

A handful of studies has comparatively evaluated various combinations of system-assigned and user-selected passwords, pronounceable passwords, and passphrases, usually with small sample sizes and student participants. Leonhard and Venkatakrishnan compared random passwords (six characters), pronounceable passwords of their own construction, and three-word passphrases drawn from Diceware in a study of 29 student participants, finding that participants had difficulty remembering the system-assigned secrets they tested [37]. In two studies using about 50 undergraduate participants each, Keith et al. found that participants with user-selected passphrases experienced more login failures due to typographical errors, but fewer failures due to memory errors, than users with passwords [27, 28]. Zviran and Haga studied 103 graduate students in a within-subjects design, where each participant used an 8-character system-assigned password, an 8-character user-selected password, and a user-selected passphrase. They found that system-assigned passwords were remembered best if they were pronouncable, but that participants preferred user-selected secrets [61]. In a within-subjects study of 15 participants, Spector and Ginzberg compared system-assigned passwords, user-selected passwords, and user-selected "pass-sentences," defined as entities with unique semantic meanings; they found that pass-sentences performed best [53]. In contrast to these studies, we use a much larger sample size and test more conditions. We focus only on system-assigned passwords and passphrases, allowing us to construct conditions with roughly similar security guarantees.

2.3.2 Passphrase Entry and Error Correction

Some researchers have proposed using error correction to permit small errors in passphrase entry, even though this may reduce security. For instance, Bard proposes using the Damerau-Levenshtein string-edit-distance metric to tolerate up to two spelling errors per word in a passphrase, as well as accepting the words of a passphrase in any order, with minimal impact on security [3]. Mehler and Skiena propose a more general password-corrective hash in which two strings differing by one edit would likely hash to the same key, successfully authenticating a user [39]. Jakobsson and Akavipat propose several error-correction techniques specific to mobile devices: allowing words in any order, permitting substitution with synonyms, and using auto-complete features for quicker entry [25]. Matsuura instead proposes a secure visual feedback mechanism to cue users when the password they have typed differs from their expectation [38]. In this paper, we evaluate participants' login attempts based on the string-edit distance and other methods that would have allowed for authentication in the presence of typing mistakes. We also test one condition in which users are explicitly told that the words of the passphrase may be entered in any order.

3. METHODOLOGY

We conducted a two-part online study of system-assigned authentication secrets, both passwords and passphrases. In the first part of the study, participants were assigned a secret, completed a survey, and were asked to recall that secret. Forty-eight hours later, participants were invited to return, log in using this secret, and complete a second survey. In this section, we give an overview of our study design, experimental conditions, and statistical analysis.

3.1 Study Overview

We recruited participants through Amazon's Mechanical Turk (MTurk) crowdsourcing service. We compensated them 55 cents for completing the first part of the study and an additional 70 cents for completing the second part. We required participants to be at least 18 years old and not to have participated in a previous study on passwords conducted by our research group. Since we tested passphrases generated with American English dictionaries, we allowed only MTurk users who lived in the United States to participate.

In part one of the experiment, we told participants, "imagine that your main email service provider has been attacked ... [and that] because of the attack, your email service provider is also changing its password rules. Instead of choosing your own password, one will be assigned to you." We informed them that they would use their secret in a few days to log in to the second part of the study, and that they should take whatever steps they would normally take to remember and protect their email password. In prior work on user-generated passwords, we observed that users created stronger passwords when presented with this scenario than when they were creating passwords simply for the purpose of a study [30, 32]. We did not tell participants not to store their secrets, nor did we otherwise mention secret storage during the first part of the study.

We next assigned participants a secret in one of 11 conditions, described in Section 3.2. After being assigned the secret, participants were required to check a box on the screen to hide the secret and then enter the secret twice, once as confirmation. They could uncheck the box to see their assigned secret again, but could not type while their secret was visible. After successfully entering and confirming their secret, participants completed a five-minute survey. This survey asked participants about their experience learning their new secret and about their actual email password. We then asked participants to enter their system-assigned secret once again. We refer to this as *part one recall* throughout the paper. After five unsuccessful attempts, participants were told their secret.

Forty-eight hours after completing the first part of the study, participants received an email through MTurk asking them to return for part two. To begin part two, participants were asked to log in using their secret. Participants could click a "Forgot Password" link to be emailed a link to retrieve their secret. Furthermore, after five incorrect attempts, we showed participants their secret. Once they had logged in, participants completed a survey about how they had remembered their secret, including whether they had written it down on paper or stored it electronically. We analyzed data for part two for only those participants who completed this part within 72 hours of being invited to return, 120 hours after completing part one. Participants who returned after more than 72 hours were still paid, but their data were excluded. This ensures that all of the participants in our analysis had viewed their secret within five days of completing part one of our study.

Mechanical Turk. As explained, we recruited participants using MTurk. Although MTurk workers tend to be younger, more educated, and more technical than the general population, they represent a significantly more diverse population than is typically used in lab studies, which often rely on college-student participants [9, 24]. Researchers have found that well-designed MTurk studies provide high-quality user data [4, 16, 21, 31, 55]. Using MTurk allows us to study a larger volume of participants in a controlled setting than would otherwise be possible. We have successfully used MTurk

to collect password data in several prior studies [30, 32, 51]. Adar has criticized MTurk studies in general, although our use of crowdsourcing to understand human behavior fits his description of an appropriate use [2].

3.2 Conditions

We assigned participants round-robin to one of 11 experimental conditions, which are summarized in Table 1. The conditions varied in the type of secret assigned to the participant. Participants were unable to modify their assigned secret or to obtain a replacement. We focused on system-assigned secrets so that we could precisely control their entropy (and their guessability), and focus on their usability.

Three conditions were variants of passwords, and eight were variants of passphrases, as defined in Section 2. Our password conditions did not use spaces. In the passphrase conditions we required that participants enter words separated by spaces and in the same order they were assigned, unless otherwise specified. Secrets in all 11 conditions were case-sensitive.

We designed two of our three password conditions and six of our eight passphrase conditions to have approximately 30 bits of entropy so that we could compare system-assigned passwords to equally strong system-assigned passphrases. This entropy value was chosen because guidelines frequently used in practice [10, 22] recommend password policies that provide an estimated 30 bits of entropy. While recent research has suggested that entropy may not be the best indicator of resilience to attack [30, 57], when all elements from a set occur with equal probability (as is the case with our system-assigned secrets), entropy maps directly to the probability that an attacker with knowledge of the password-generation algorithm will succesfully guess a password.

3.2.1 Password conditions

Three of our conditions focused on passwords.

- **pw-length5**: Participants were assigned a five-character password, where each character is chosen randomly from a dictionary of 64 characters, including lowercase letters, uppercase letters, digits, and symbols. We removed characters that could easily be confused with other characters, e.g., both the letter "O" and the digit "0." A full list of characters in this dictionary is shown in Appendix D.1. This password space has 30 bits of entropy.

- **pw-pronounce**: Participants were assigned an eight-character password likely to be pronounceable by an English speaker. To generate these passwords, we used an implementation[5] of an algorithm originally proposed by Gasser [19] and later adopted as a NIST standard [41]. Prior work has identified a flaw in this scheme: certain passwords are chosen with high probability since the probability of a syllable occuring in a password mirrors its relative frequency in English [36]. To overcome this, we generated the full list of eight-character pronounceable passwords without duplicates ($\approx$ 1.2 billion) and assigned each password on this list with equal probability, resulting in 30.2 bits of entropy.

- **pw-length6**: Participants were assigned a six-character password, where characters are chosen as in the pw-length5 condition. The extra character makes passwords in this condition have 36 bits of entropy. This condition helps determine how the length of randomly generated passwords affects usability.

[5] `http://www.adel.nursat.kz/apg/` (visited 5/2012)

3.2.2 Passphrase conditions

We tested eight variations on passphrases. We generated dictionaries (fully specified in Appendix D) for all passphrase conditions using word-frequency data[6] with part-of-speech (e.g., noun, verb) tags from the Corpus of Contemporary American English (COCA) [14]. This list ranks the most common words in the 425-million-word COCA based on the number of times they appear and their diffusion throughout different sources. So that our dictionaries would contain only well-known words, we chose the N most common words matching particular criteria for each dictionary. For instance, a dictionary of 181 nouns would contain the 181 most common nouns from COCA. We selected word lists of particular sizes so that different conditions would each have 30 bits of entropy. However, we later discovered that word lists not restricted to a particular part of speech contained duplicate words. For instance, "to" was present on the list as both an infinitive marker and as a preposition. Thus, the actual passphrase entropies in the next three conditions, intended to be 30 bits, were as low as 29.3 bits.[7]

- **pp-small**: Participants were assigned four words, randomly selected with replacement from a 181-word dictionary.

- **pp-med-unorder**: Participants were assigned four words, randomly selected with replacement from a 401-word dictionary. Unlike all other conditions, participants could enter the words in their passphrase in any order.

- **pp-large-3word**: Participants were assigned three words, randomly selected with replacement from a 1,024-word dictionary.

The next two conditions are similar to the pp-small condition, except they use larger dictionaries of the most common words in order to test whether the size of the dictionary has a measurable impact on usability:

- **pp-medium**: Participants were assigned four words, randomly selected with replacement from the 401-word dictionary used in the pp-med-unorder condition. These passphrases present 33.9 bits of entropy.

- **pp-large**: Participants were assigned four words, randomly selected with replacement from the 1,024-word dictionary used in the pp-large-3word condition. These passphrases present 39.2 bits of entropy.

We also tested whether passphrases that followed certain part-of-speech patterns aid memorability. The next three conditions use passphrases with 30 bits of entropy.

- **pp-sentence**: Participants were assigned passphrases of the form "noun verb adjective noun," where nouns, verbs, and adjectives are chosen from separate 181-word dictionaries. So that it would make sense for the verb to be followed by a noun, the verb dictionary contained only verbs whose entry in The Free Dictionary[8] listed at least one transitive definition. Since all nouns but one were singular, we manually conjugated all verbs to agree with a singular subject. Although these passphrases were unlikely to make semantic sense due to the random selection of words, they might resemble English sentences.

[6] `http://www.wordfrequency.info/top5000.asp` (visited 5/2012)

[7] All entropies were calculated using Shannon's formula on the frequency distribution of unique words [50].

[8] `http://www.thefreedictionary.com/` (visited 5/2012)

Condition name	Entropy (bits)	Length	Dictionary size	Examples	
Password Conditions					
pw-length5	30	5 characters	64 characters	@J#8x	*2LxG
pw-pronounce	30.2	8 characters	190 syllables	tufritvi	vadasabi
pw-length6	36	6 characters	64 characters	R6wy$_	cW@.*H
Passphrase Conditions					
pp-small	29.4	4 words	181 words	one between high tell	try there three come
pp-med-unorder	29.3	4 words (any order OK)	401 words	remember million state understand	help any country our
pp-large-3word	29.4	3 words	1,024 words	own decide some	feeling right reflect
pp-nouns	30	4 nouns	181 nouns	sense child reason paper	death effect girl model
pp-nouns-instr	30	4 nouns (w/ instructions)	181 nouns	phone star record right	case area interest situation
pp-sentence	30	4 words (N-V-Adj-N)	181 words each	end determines red drug	plan builds sure power
pp-medium	33.9	4 words	401 words	also that research must	room four face after
pp-large	39.2	4 words	1,024 words	because strategy cover us	pull somebody white next

Table 1: A summary of experimental conditions, with data about their characteristics and example secrets assigned to participants.

- **pp-nouns**: Participants were assigned four nouns, randomly sampled with replacement from a dictionary containing the 181 most common nouns.

- **pp-nouns-instr**: The condition is identical to pp-nouns, except that we gave the participant specific instructions for memorizing the passphrase. The instructions asked participants to "try to imagine a scene that includes all of the words in your password phrase. This will help you to remember it more easily. Research has found that the more bizarre, unusual, and exaggerated you make your scene, the easier it will be to remember. So, take a moment to construct your scene, and think about it whenever you need to enter your password." This specific instruction mimics the example from the xkcd comic in Appendix A.

3.3 Statistical Testing

Our statistical tests use a significance level of $\alpha = .05$. For each comparison, we first ran an omnibus test across all conditions. We used Kruskal-Wallis (indicated KW), an analogue of ANOVA that does not assume normality, for omnibus tests on quantitative data and χ^2 on categorical. If the omnibus test showed significance, we performed pairwise tests with Holm-Bonferroni correction (indicated HC). We used Mann-Whitney U (indicated MW) for pairwise comparisons of quantitative data and Fisher's Exact Test (indicated FET) for pairwise categorical comparisons.[9]

In our pairwise tests, we compared a subset of all possible pairs of conditions. All eight of our 30-bit conditions are compared to each other. We also compare pp-medium with pp-med-unorder, because they both use a 401-word dictionary; and pp-large with pp-large-3word since both use 1,024-word dictionaries. We compare pw-length5 with pw-length6 as the latter uses longer passwords, but they are otherwise identical. Finally, we compare pp-medium with pw-length6 to compare a password and a passphrase condition with higher entropy.

In addition to looking at conditions independently, we sometimes combine a subset of our password conditions and a subset of our passphrase conditions to compare larger sample sizes of passwords and equivalent-entropy passphrases. The *combined passwords participants* comprise our two 30-bit password conditions, pw-length5 and pw-pronounce. The *combined passphrases participants* comprise our 30-bit passphrase conditions that use 181-word dictionaries: pp-small, pp-sentence, pp-nouns, and pp-nouns-instr.

[9]When using the χ^2 test, some cell counts were less than 5, but we ensured that Cochran's conditions were satisfied: no cell had count zero, and more than 80% of cells had counts of at least 5.

4. RESULTS

In this section, we present the results of our study. We begin by discussing participant demographics in Section 4.1. We then look at drop-out rates per condition in Section 4.2, as higher drop-out rates may indicate participants are struggling more in those conditions. In Section 4.3, we define how we tracked whether participants stored their assigned secret. We examine how well participants were able to enter their secrets immediately upon assignment in Section 4.4. Section 4.5 discusses part-one recall rates, and Section 4.6 part-two recall rates. We further investigate usability by examining user sentiment in Section 4.8. In Section 5, we analyze the errors participants made, and the degree to which automated error correction would have helped.

4.1 Demographics

2,689 participants began our study in February and early March 2012. 2,294 completed the first part of our study and 1,562 returned for the second part within three days of being sent an email invitation two days after completing the first part. An additional 88 participants returned for the second part between three and 42 days after completing the first part; we do not include them in our analysis. Of the participants who returned within three days, 1,476 completed the second part of our study. With the exception of our discussion of drop-out rates in Section 4.2, we focus on these 1,476 participants throughout our analysis. The number of participants in each condition is shown in Table 2.

Of the 1,468 participants who reported gender, 51.9% reported being female and 47.6% reported being male. The mean age was 31 years, while the median was 28. The standard deviation was 11.2, and our oldest participant reported being 74 years old.

Of the 1,464 participants who reported their highest academic degree, 653 reported having at least a bachelor's degree. Participants were asked whether they had degrees or jobs in "computer science, computer engineering, information technology, or a related field." Of the 1,460 who answered, 263 answered in the affirmative. We found no statistically significant differences between our conditions in reported gender, age, or background and education.

Because using a keyboard on a mobile phone could impact a participant's ability to enter his or her secret, we examined participants' user-agent strings. For example, if a user-agent string contains "iPhone," that is strong evidence that the participant is taking

	Started	Finished part one	Returned	Finished part two
pw-length5	342	90%	59%	56%
pw-pronounce	342	90%	58%	55%
combined password	684	90%	59%	55%
pp-small	189	84%	59%	59%
pp-nouns	340	81%	59%	55%
pp-nouns-instr	342	87%	59%	56%
pp-sentence	190	79%	54%	50%
combined passphrase	1061	83%	58%	55%
pp-med-unorder	187	85%	58%	57%
pp-large-3word	188	88%	61%	55%
pp-medium	190	84%	57%	53%
pw-length6	190	83%	58%	55%
pp-large	189	82%	54%	53%
total	2689	85%	58%	55%

Table 2: The number of participants who began the study in each condition, and the percentage who continued through the steps of the study. A participant counts as having returned for the second part of the study if he or she returned within three days of being invited. The analysis in this paper focuses on participants who completed the second part of the study.

success on first entry – *omnibus* χ^2_{10}=28.026, p=.002		
pw-pronounce (90.4%)	pp-nouns (73.4%)	HC FET, p=.001
	pp-nouns-instr (71.2%)	HC FET, p<.001
combined pw (84.9%)	combined pp (74.5%)	χ^2_1=14.233, p<.001
success on first entry (no-storage) – *omnibus* χ^2_{10}=27.021, p=.003		
pw-pronounce (92.3%)	pp-nouns-instr (67.1%)	HC FET, p=.027
attempts needed – *omnibus* KW χ^2_{10}=28.573, p=.001		
pw-pronounce (1.14)	pp-nouns (1.45)	U=14526, p<.001
	pp-nouns-instr (1.50)	U=14367.5, p<.001
	pp-sentence (1.33)	U=7473, p=.025
combined pw (1.23)	combined pp (1.41)	KW χ^2_1=14.979, p<.001
attempts needed (no-storage) – *omnibus* KW χ^2_{10}=28.888, p=.001		
pw-length5 (1.24)	pw-length6 (2.55)	U=274.5, p=.037
pw-pronounce (1.08)	pp-nouns-instr (1.53)	U=1350, p=.025

Table 3: Statistically significant results for secret entry immediately upon assignment. Pairwise tests for attempts needed are Holm-Bonferonni-corrected Mann-Whitney U.

the study from an iPhone. Only 25 participants show evidence of this,[10] and there were no more than four per condition.

4.2 Study Dropouts

Of the 2,689 participants who started our study, 2,294 finished the first part; 1,562 participants returned within three days of receiving our email invitation to complete the second part of the study, and 1,476 of these completed the second part. These statistics, broken down by condition, are shown in Table 2.

The proportion of participants who completed the first part of the study varied by condition (χ^2_{10}=28.288, p=.002), with participants in the two 30-bit password conditions, surprisingly, most likely to finish. Completion rates for the first part ranged from 78.9% for pp-sentence to 90.4% for pw-pronounce. Significantly more participants finished the first day in pw-pronounce than in pp-nouns (HC FET, p=.020) and pp-sentence (HC FET, p=.011). More also completed the first day in pw-length5 than in pp-sentence (HC FET, p=.035). Combined password participants were more likely to finish than combined passphrase participants (χ^2_1=14.768, p<.001).

The proportion of participants who returned for the second part of the study within five days after completing the first did not vary significantly by condition (χ^2_{10}=6.759, p=.748), and neither did the proportion of these who finished the second part (χ^2_{10}=15.956, p =.101). We also saw no significant difference between combined password participants and combined passphrase participants for returning within five days (χ^2_1=3.423, p=.064) or finishing the second day (χ^2_1=0.015, p=.901).

4.3 Storage and No-Storage Participnts

During the second part of the study, we asked participants if they wrote down their secrets, either on paper or electronically, reassuring them that their compensation would not be affected by their response. We consider a participant not to have stored his or her secret if the participant affirms not writing down the secret, and we do not detect that he or she has pasted or autofilled the secret. We label these participants, 410 of our 1,476 total (27.8%), as *no-storage*; other participants we call *storage* participants. The proportion of no-storage participants is low across conditions; it does not vary significantly by condition (χ^2_{10}=17.351, p=.067), nor is it significantly different between the combined password participants and the combined passphrase participants (χ^2_1=2.444, p=.118).

The no-storage participants are most relevant when evaluating the memorability of secrets. However, as users can and do store secrets for their real accounts, the behavior of storage participants also provides useful insights. We do not know when storage participants stored their secrets (e.g., before entering them for the first time, or after completing part one). Hence, while we divide participants based on whether they eventually stored their secret, we cannot separately analyze them based on when the secret was stored.

In addition to analyzing data for all participants who did not drop out of the study, we conducted similar analyses separately for storage and no-storage participants. In some cases, differences between conditions that were statistically significant when looking at all participants are no longer significant when looking only at no-storage participants. However, this may be due in part to the small number of no-storage participants. We revisit the results of these separate analyses in the remainder of this section and in Section 5.

4.4 Assignment

After receiving instructions, our participants were assigned a secret (either a password or a passphrase) and immediately asked to enter it. They could toggle between being able to view the secret and being able to enter it. This was intended to ensure that participants observed and were able to type their secret. We measure the number of attempts needed to enter the secret successfully, and the fraction of participants who correctly entered their secret on the first try. Significant results of statistical tests are shown in Table 3.

Overall, participants needed an average of 1.3 attempts to enter their secret and 78.7% entered it successfully on the first try. For both metrics, there was a significant difference across conditions. pw-pronounce secrets needed fewer attempts and were more likely to be entered on the first try than pp-nouns-instr and pp-nouns, and also needed fewer attempts than pp-sentence. In aggregate, combined password participants needed fewer attempts and were more successful in entering their secret on the first attempt than combined passphrase participants.

Similar relationships hold for no-storage participants: pw-pronounce needed fewer attempts and was more successful on the first

[10]We searched the user-agent strings for: Android, iPhone, iPod, iPad, mobile, RIM Tablet, BlackBerry, Opera Mini, Windows Phone, SymbianOS, Opera Mobi, nook, Windows CE, smartphone, webOS, BREW.

attempt than pp-nouns-instr. No-storage participants also show difference based on password length, with pw-length5 requiring fewer attempts than pw-length6.

4.5 Part-One Recall

We asked participants to recall their secret after completing a brief survey. Participants who could not recall their secret after five attempts were shown the secret on the screen. The vast majority of participants in each condition succeeded in entering their secret within five attempts, ranging from 92.5% in pp-med-unorder to 99.5% in pw-length5 and pw-pronounce. The significant results of our statistical tests are in Table 4.

Among those who entered the secret within five attempts, there is a significant difference in the number of attempts needed; combined password participants took significantly fewer attempts than combined passphrase participants, but pairwise tests do not reveal any significant differences.

The proportion of participants who correctly entered their secret on the first try varied significantly across conditions. A larger proportion in pw-length5 and pw-pronounce entered their password correctly on the first attempt than in pp-small or pp-med-unorder. Combined password participants outperformed combined passphrase participants.

Among no-storage participants, significantly more in pw-pronounce entered their secret correctly on the first try than in pp-large-3word. Looking at participants who entered their secret within five attempts, we see omnibus significance between conditions, but pairwise tests reveal no significant differences.

We measured the time between the first and last keystroke on the first correct entry for participants whom we did not detect pasting or autofilling their secrets, and who entered the secret within five attempts. Combined passphrase participants had a median time of 7 seconds, significantly more than combined password participants, with a median of 3 seconds. pw-length5 and pw-pronounce each performed significantly better than all of the 30-bit passphrase conditions, and pw-length6 performed better than pp-medium. pp-small and pp-large-3word performed significantly better than pp-nouns, pp-nouns-instr, and pp-sentence; and pp-large-3word also outperformed pp-large. Looking only at no-storage participants, pw-pronounce performed better than any other 30-bit condition except pw-length5. pw-length5 and pp-small both outperformed pp-nouns, pp-nouns-instr, and pp-sentence.

4.6 Part-Two Recall

Forty-eight hours after finishing the first part of our study, we invited participants to return for the second part. Our analysis includes the participants who returned within 72 hours of being invited and completed both parts of the study. We find that a majority in each condition wrote down their secrets, and nearly half that did not store their secret clicked on the "Forgot Password" link.

Upon returning, a participant was asked to recall his or her secret. Five incorrect entries resulted in the secret being shown on screen. How participants fared in entering their secrets here is shown in Table 5. Our statistical tests are shown in Table 6.

Returning participants could click a link to be emailed a link to their secret. 48.8% of no-storage participants used this feature. The proportion does not vary significantly by condition (χ^2_{10}=11.992, p =.286), nor between combined passphrase participants and combined password participants (χ^2_1=1.764, p=.184). 210 of our 1,476 participants did not use the email reminder or store their secrets. Across all conditions, out of the 354 participants who used the email reminder, 197 (55.6%) made no attempt to recall their secrets and 87 (24.6%) made only one attempt before using the reminder.

success on first entry – *omnibus* χ^2_{10}=34.936, p<.001		
pw-length5 (94.8%)	pp-small (81.1%)	HC FET, p=.008
	pp-med-unorder (80.2%)	HC FET, p=.007
pw-pronounce (94.7%)	pp-small (81.1%)	HC FET, p=.009
	pp-med-unorder (80.2%)	HC FET, p=.007
combined pw (94.7%)	combined pp (87.2%)	χ^2_1=13.867, p<.001
success on first entry (no-storage) – *omnibus* χ^2_{10}=26.558, p=.003		
pw-pronounce (94.2%)	pp-large-3word (64.3%)	HC FET, p=.033
attempts needed – *omnibus* KW χ^2_{10}=19.122, p=.039		
combined pw (1.08)	combined pp (1.16)	KW χ^2_1=8.066, p=.005
attempts needed (no-storage) – *omnibus* KW χ^2_{10}=20.002, p=.029		
secret entry time – *omnibus* KW χ^2_{10}=329.817, p<.001		
pw-pronounce (3.1)	pp-small (5.3)	U=3296, p<.001
	pp-nouns (7.4)	U=3187.5, p<.001
	pp-nouns-instr (7.4)	U=2833, p<.001
	pp-sentence (7.7)	U=1310.5, p<.001
	pp-large-3word (4.7)	U=3626, p<.001
	pp-med-unorder (6.1)	U=2548.5, p<.001
pw-length5 (3.4)	pp-small (5.3)	U=3962, p<.001
	pp-nouns (7.4)	U=4126, p<.001
	pp-nouns-instr (7.4)	U=3652.5, p<.001
	pp-sentence (7.7)	U=1792, p<.001
	pp-large-3word (4.7)	U=4223.5, p<.001
	pp-med-unorder (6.1)	U=3141.5, p<.001
	pw-length6 (4.2)	U=4857.5, p=.039
pw-length6 (4.2)	pp-medium (6.5)	U=1933, p<.001
pp-small (5.3)	pp-nouns (7.4)	U=4717.5, p=.009
	pp-nouns-instr (7.4)	U=4328.5, p<.001
	pp-sentence (7.7)	U=2212.5, p=.002
pp-large-3word (4.7)	pp-nouns (7.4)	U=4106.5, p=.001
	pp-nouns-instr (7.4)	U=3809, p<.001
	pp-sentence (7.7)	U=1914, p<.001
	pp-large (7.4)	U=2101.5, p=.001
pp-med-unorder (6.1)	pp-nouns-instr (7.4)	U=4886, p=.015
combined pw (3.1)	combined pp (7.0)	KW χ^2_1=249.884, p<.001
secret entry time (no-storage) – *omnibus* KW χ^2_{10}=130.86, p<0		
pw-pronounce (2.6)	pp-small (3.9)	U=792, p=.017
	pp-nouns (7.6)	U=200, p<.001
	pp-nouns-instr (7.4)	U=226, p<.001
	pp-sentence (6.9)	U=98, p<.001
	pp-large-3word (4.2)	U=258, p=.008
	pp-med-unorder (5.4)	U=144, p<.001
pw-length5 (3.4)	pp-nouns (7.6)	U=389, p<.001
	pp-nouns-instr (7.4)	U=425, p<.001
	pp-sentence (6.9)	U=189, p<.001
pp-small (3.9)	pp-nouns (7.6)	U=214, p=.002
	pp-nouns-instr (7.4)	U=240, p<.001
	pp-sentence (6.9)	U=93, p=.001
combined pw (2.8)	combined pp (6.9)	KW χ^2_1=89.758, p<.001

Table 4: Statistically significant results for secret recall after completing the survey in part one of the study. Times are shown as median seconds. Attempts needed are shown for participants who succeeded in entering their secret within five attempts. All pairwise tests for secret-entry time are Holm-Bonferonni-corrected Mann-Whitney U.

We consider a participant to have succeeded in recalling his or her secret in the second part of our study if he or she entered the secret within five attempts without needing to be reminded of it. Overall, 74.7% of our participants were successful, including 48.5% of no-storage participants and 84.7% of storage participants. There were no significant differences between conditions, nor between combined password and combined passphrase participants, in any of these groups.

For all participants, and for just no-storage participants, we see no significant difference in how many participants succeed on the

	No-storage			Storage		
	Participants	Login in five tries	Login on first try	Participants	Login in five tries	Login on first try
pw-length5	46	65%	57%	145	86%	81%
pw-pronounce	52	52%	48%	135	83%	77%
combined password	98	58%	52%	280	85%	79%
pp-small	26	42%	38%	85	86%	76%
pp-nouns	57	47%	42%	131	84%	76%
pp-nouns-instr	70	51%	37%	121	83%	78%
pp-sentence	27	41%	37%	67	87%	76%
combined passphrase	180	47%	39%	404	85%	77%
pp-med-unorder	25	36%	36%	81	80%	70%
pp-large-3word	28	57%	50%	75	85%	75%
pp-medium	34	35%	32%	67	81%	73%
pw-length6	20	45%	40%	84	92%	82%
pp-large	25	44%	40%	75	85%	73%
total	410	49%	42%	1066	85%	77%

Table 5: Successful logins in the second part of the study for no-storage and storage participants. Participants are considered not to have been successful in five tries if they either entered their secret unsuccessfully five times or requested to have their secret emailed to them.

first attempt between conditions.

For each participant on his or her first attempt at secret entry, we calculated the edit distance between what was entered and the assigned secret. We use the Damerau-Levenshtein edit-distance metric, which is the minimum number of insertions, deletions, substitutions, and adjacent transpositions required to transform one string into another. The mean edit distance on the first attempt is shown, per condition, in Table 7. The mean edit distance was less than one for each of the password conditions, and for passphrases it ranged between 1.12 for pp-large-3word and 2.96 for pp-nouns-instr. The median for each condition was zero, and the edit distance did not vary significantly between conditions, either for all or just no-storage participants.

Looking at successful no-storage participants in the pp-med-unorder condition, six of nine entered the password in the same order as it was assigned. Overall 68 out of 74 participants in pp-med-unorder entered the passphrase in the same order as it was assigned.

Another metric for usability we examined was the use of *deletes* during secret entry. A delete may indicate a participant changing his or her mind about a secret while entering it. We counted each instance of one or more characters being removed from the secret-entry field as a single delete and recorded the number per secret-entry attempt for each participant. Deletions per condition are shown in Table 7. Looking only at participants who succeeded in entering their secret on the first try in the second part of the study, we find the conditions did differ significantly in the number of deletions. The mean for each password condition was less than one, while for passphrase conditions it ranged from 1.76 for pp-medium to 3.78 for pp-sentence. pw-length5 had significantly fewer deletions than any 30-bit passphrase condition except pp-small, and pw-pronounce had significantly fewer deletions than pp-nouns-instr. Looking only at no-storage participants, the omnibus test shows significance, but pairwise comparisons do not.

Another usability metric is *login* time, the total time a participant took to enter his or her secret, measured from the participant's first arrival at the secret-entry screen until the end of the participant's last visit to that screen. This includes anything between

success – *omnibus* χ^2_{10}=15.584, p=.112

comb. pw (77.8%)	comb. pp (73.1%)	χ^2_1=2.413, p=.120

success (no-storage) – *omnibus* χ^2_{10}=11.786, p=.3

comb. pw (58.2%)	comb. pp (47.2%)	χ^2_1=2.618, p=.106

success (storage) – *omnibus* χ^2_{10}=6.341, p=.786

comb. pw (84.6%)	comb. pp (84.7%)	χ^2_1=.011, p=.917

attempts for success. participants – *omnibus* KW χ^2_{10}=4.376, p=.929

comb. pw (1.112)	comb. pp (1.124)	KW χ^2_1=2.118, p=.146

successful first try – *omnibus* χ^2_{10}=4.774, p=.906
successful first try (no-storage) – *omnibus* χ^2_{10}=9.797, p=.458

edit distance – *omnibus* KW χ^2_{10}=12.579, p=.248
edit distance (no storage) – *omnibus* KW χ^2_{10}=10.407, p=.406

deletions – *omnibus* KW χ^2_{10}=36.614, p<.001

pw-length5 (0.17)	pp-nouns (3.02)	U=9660.5, p=.002
	pp-nouns-ins (2.72)	U=9078.5, p=.001
	pp-sentence (3.78)	U=4754, p=.017
	pp-lg-3word (2.67)	U=5536, p=.010
	pp-med-unorder (2.67)	U=5372, p=.008
pw-pronounce (0.7)	pp-nouns-ins (2.72)	U=9327.5, p=.025

deletions (no storage) – *omnibus* KW χ^2_{10}=18.324, p=.05

login time – *omnibus* KW χ^2_{10}=36.259, p<.001

pw-pronounce (25)	pp-nouns (35)	U=13716.5, p=.018

login time (no-storage) – *omnibus* KW χ^2_{10}=15.79, p=.106

secret entry time – *omnibus* KW χ^2_{10}=204.592, p<.001

pw-length5 (4.0)	pw-length6 (5.5)	U=3931.5, p=.021
	pp-small (5.3)	U=4063.5, p=.017
	pp-nouns (8.4)	U=4256, p<.001
	pp-nouns-ins (8.6)	U=4736, p<.001
	pp-sentence (9.0)	U=2109.5, p<.001
	pp-lg-3word (5.1)	U=4152.5, p=.016
	pp-med-unorder (6.5)	U=3193, p<.001
pw-pronounce (3.3)	pp-small (5.3)	U=3765, p=.001
	pp-nouns (8.4)	U=3894.5, p<.001
	pp-nouns-ins (8.6)	U=4363.5, p<.001
	pp-sentence (9.0)	U=1899.5, p<.001
	pp-lg-3word (5.1)	U=3809, p<.001
	pp-med-unorder (6.5)	U=3002.5, p<.001
pp-small (5.3)	pp-nouns (8.4)	U=3675, p<.001
	pp-nouns-ins (8.6)	U=4078, p=.002
	pp-sentence (9.0)	U=1842, p=.003
pp-lg-3word (5.1)	pp-nouns (8.4)	U=4045, p=.002
	pp-nouns-ins (8.6)	U=4438.5, p=.013
	pp-sentence (9.0)	U=1982, p=.009
	pp-large (7.8)	U=2274, p=.013
comb. pw (3.6)	comb. pp (7.9)	KW χ^2_1=148.919, p<.001

secret entry time (no storage) – *omnibus* KW χ^2_{10}=91.124, p<.001

pw-length5 (3.7)	pp-nouns (7.6)	U=457, p<.001
	pp-nouns-ins (7.2)	U=687, p<.001
	pp-sentence (6.0)	U=230, p=.007
pw-pronounce (2.7)	pp-nouns (7.6)	U=380, p<.001
	pp-nouns-ins (7.2)	U=577, p<.001
	pp-sentence (6.0)	U=186, p<.001
	pp-med-unorder (5.1)	U=270, p=.009
pp-small (4.2)	pp-nouns (7.6)	U=188, p<.001
	pp-nouns-ins (7.2)	U=327, p=.005
	pp-sentence (6.0)	U=109, p=.031
pp-lg-3word (4.1)	pp-nouns (7.6)	U=319, p=.008
comb. pw (3.1)	comb. pp (6.8)	KW χ^2_1=55.932, p<.001

Table 6: Selected results for secret recall after returning for the second part of the study. Success indicates the percentage of participants who entered the secret in five attempts without an email reminder. Times are shown as median seconds. Deletions are for participants who correctly entered their secret on the first try. All pairwise tests are Holm-Bonferonni-corrected Mann-Whitney U.

		All participants				No-storage			
	Mean length (chars)	Med. entry time (s)	Med. login time (s)	Mean deletions	Mean edit dist.	Med. entry time (s)	Med. login time (s)	Mean deletions	Mean edit dist.
pw-length5	5.0	4.0	27.5	0.2	0.9	3.7	32.5	0.2	1.5
pw-pronounce	8.0	3.3	25.0	0.7	0.9	2.7	26.0	0.8	1.1
combined password	6.5	3.6	26.0	0.4	0.9	3.1	27.5	0.5	1.3
pp-small	18.3	5.3	26.0	2.0	1.7	4.2	31.0	0.9	4.4
pp-nouns	24.2	8.4	35.0	3.0	2.5	7.6	35.0	3.9	3.2
pp-nouns-instr	24.7	8.6	34.0	2.7	3.0	7.2	31.0	2.3	4.0
pp-sentence	25.5	9.0	34.0	3.8	2.4	6.0	31.0	5.6	2.1
combined passphrase	23.4	7.9	33.0	2.8	2.5	6.8	32.0	3.2	3.5
pp-med-u	21.3	6.5	36.0	2.7	2.5	5.1	28.5	2.6	1.9
pp-large-3word	18.4	5.1	27.0	2.7	1.1	4.1	23.5	1.7	2.5
pp-medium	21.2	6.6	35.0	1.8	2.3	6.6	44.5	3.3	5.4
pw-length6	6.0	5.5	24.0	0.4	0.8	5.1	44.5	0.2	1.6
pp-large	24.1	7.8	34.5	3.5	2.5	7.0	35.0	3.8	3.7
total	17.2	6.0	31.0	2.0	1.9	5.6	32.0	2.3	2.9

Table 7: Length, entry time, login time, number of deletions and edit distance for each condition. Entry time is the median secret-entry time, first to last keystroke, in the second part of our study, for participants who did not paste or autofill their secrets, and who entered them within five attempts. Login time is the median time between a participant first being shown the second-part recall screen and leaving that page for the final time. Deletions are counted for those participants who succeeded at their first secret-entry attempt in the second part of the study. Edit distance is computed between the actual secret and what was entered, for all participants in their first attempt in the second part of the study.

visits to that screen, like using the secret-reminder feature. Login times by condition are shown in Table 7. Login time varies significantly across conditions, with pw-pronounce performing significantly better than pp-nouns. Login time does not vary significantly by condition for no-storage participants.

We next examined the time between the first and the last keystroke on the initial correct entry for participants who neither pasted nor autofilled their secrets in the second part of our study, and entered their secret within five attempts. pw-length5 and pw-pronounce each performed significantly better than most of the passphrase conditions, and pp-small and pp-large-3word outperformed most other passphrase conditions; these and other comparisons are shown in Table 6, and login times are shown in Table 7. Similar relationships hold for no-storage participants: pw-length5, pw-pronounce, and pp-small outperform pp-nouns, pp-nouns-instr, and pp-sentence; pw-pronounce also performed better than pp-med-unorder, and pp-large-3word better than pp-nouns.

4.7 Storage Behavior

We examined how participants stored and protected the secrets used in this study, as well as how they reported storing their real email passwords. As indicated in Section 4.3, we assume a participant has stored his or her secret unless that participant explicitly states he or she has not written it down or otherwise stored it, and has not pasted or autofilled the secret. 72% of participants stored their secret. Of these, 48.3% indicated writing their secret down on paper and 43.6% reported storing it electronically; 23% pasted their secret. A single participant may have done more than one

annoying – *omnibus* χ^2_{10}=30.116, p=.001		
pw-length6 (61.5%)	pp-medium (33.7%)	HC FET, p=.003
difficult – *omnibus* χ^2_{10}=66.583, p<.001		
pp-large (18.0%)	pp-large-3word (1.9%)	HC FET, p=.002
all other 30-bit (14.4 - 25.0%)		HC FET, p<.023
pw-length6 (44.2%)	pp-medium (16.8%)	HC FET, p=.001
	pw-length5 (22.0%)	HC FET, p=.003
fun – *omnibus* χ^2_{10}=43.433, p<.001		
pp-nouns-instr (25.7%)	pp-small (9.0%)	HC FET, p=.014
	pp-large-3word (8.7%)	HC FET, p=.012
	pw-length5 (9.4%)	HC FET, p=.001
pp-medium (20.8%)	pw-length6 (4.8%)	HC FET, p=.018

Table 8: Statistically significant results for user sentiment.

of the above. Storage rate was not significantly different between conditions (χ^2_{10}=17.351, p=.067). This result matches Zviran and Haga's [62]; they found, surprisingly, that "difficulty recalling a password or writing it down is not related to password's length."

We asked participants "If you wrote down or stored your password for this study, how is it protected (choose all that apply)?" Of our 1,066 storage participants, 21.9% did nothing to protect their passwords. 26.7% said they stored it on a computer or device used only by themselves, the most popular response. 24.5% stored the password in a room or office used only by that participant.

We also asked our participants about their real email passwords. 308 indicated referring to a written-down or stored password when logging in with their real email password, and 1,168 did not. We also asked if they had ever stored their real email password. 768 participants indicated never writing down their real email password, while 373 did so on paper and 430 electronically. This 52.6% of participants who did not store their real passwords is a significantly larger proportion than the 32.8% who indicated not storing their study secret (χ^2_1=114.287, p<.001).

4.8 User Sentiment

In the first part of the study, we asked participants to indicate their agreement, from "strongly disagree" to "strongly agree," with the statements "learning my password was [fun/difficult/annoying]." We classify participants as either agreeing ("agree" or "strongly agree") or not agreeing with each statement. An overview of results is shown in Figure 1, with detailed statistical results in Table 8.

We see a significant difference in annoyance, fun, and difficulty memorizing across conditions. Pairwise tests show that pw-length6 was substantially more annoying, difficult to learn, and less fun than pp-medium, and was also more difficult than pw-length5. pp-large, and all 30-bit conditions, were more difficult to memorize than pp-large-3word. Finally, pp-nouns-instr was more fun by a wide margin than pp-small, pp-large-3word, or pw-length5.

Comparing our combined password and combined passphrase participants, we see no significant difference in agreement that memorizing the secrets was annoying (χ^2_1=0.219, p=.639), difficult (χ^2_1 =0.022, p=.882), or fun (χ^2_1=1.65, p=.199).

Looking only at no-storage participants, we see no significant differences in annoyance (χ^2_{10}=10.952, p=.361), and omnibus difference in difficulty (χ^2_{10}=23.317, p=.01) and fun (χ^2_{10}=23.998, p =0.008) but no pairwise significance for either.

5. ERROR ANALYSIS RESULTS

Examining factors that lead to user error can help us understand why passphrases were less successful than we anticipated, and can inform research on improving their performance. In addition, passphrases (and to a lesser extent pronounceable passwords) offer sev-

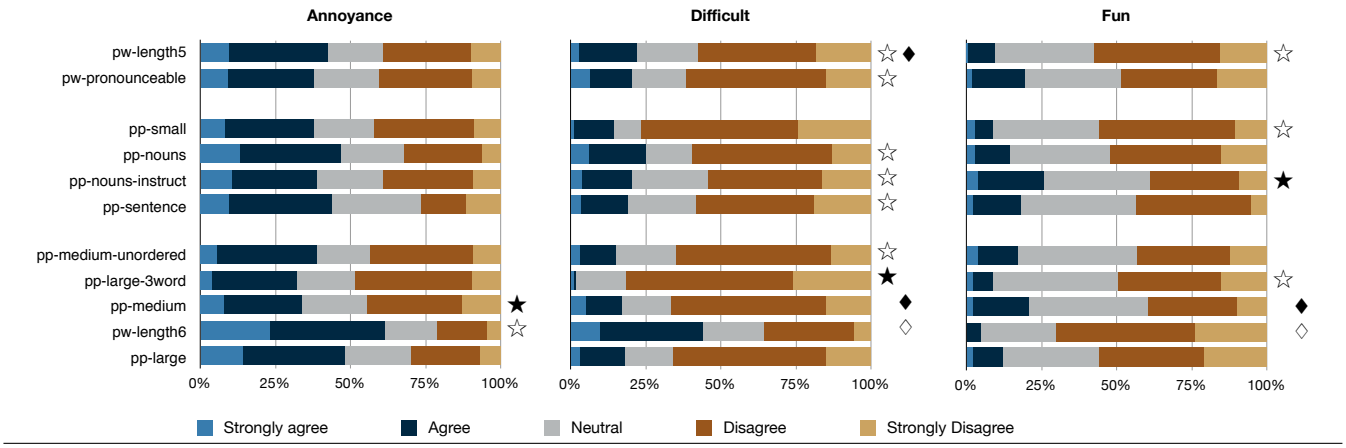

Figure 1: Likert response data on annoyance, difficulty, and fun. Data significance is shown in symbol groups; a condition marked with a solid symbol performs significantly better than conditions marked with the same symbol drawn as an outline.

eral opportunities for automatic error detection and correction, which may be able to improve usability without loss of entropy.

The overall results of our error analysis are shown in Table 9. This table displays the percentage of subjects who correctly entered their secret, both with no correction and adjusted for the use of different error-correcting mechanisms, as discussed below.

5.1 Length

We hypothesized that longer secrets (in characters) lead to more typing errors. Table 7 shows the mean length of secrets, per condition. For passphrase conditions, in which secrets are generally longer than in our password conditions, we find that longer passphrases reduce the likelihood of authentication success at assignment, but not thereafter. For all passphrase participants, we used logistic regression on passphrase length with an outcome of first-attempt success at assignment, and found a significant relationship (p=.003). The shortest passphrase condition, pp-small, had a mean length of 18.3 characters and a first-try success rate of 81% at assignment. By contrast, the longest condition, pp-sentence, had a mean of 25.5 characters and a first-try success rate of only 76%.

The same analysis for first-try accuracy for recall in part one (excluding participants who pasted or auto-filled their passphrase) and part two (excluding storage participants) found that length was not a significant factor in either case (p>.375). We also found no relationship between length and overall rate of successful login for part two (within five attempts, without using the reminder, p=.406).

5.2 Ignoring Spaces and Capitalization

We required participants to enter their secret exactly as we showed it to them, including spaces and capitalization. In general, however, passphrase dictionaries can be designed to be case-insensitive and unambiguous, even when spaces are removed. In addition, our pw-pronounce condition did not include uppercase. In such cases, removing spaces and ignoring case when checking input passwords can potentially improve usability with no cost to security. We examine how our passphrase and pw-pronounce participants would have performed had we ignored spaces and capitalization. We find that while error correction provides a small benefit, it does not cause passphrase performance to improve relative to passwords.

As shown in Table 9, ignoring case and spaces improves first-attempt accuracy for every passphrase condition as well as pw-pronounce, but has minimal impact on overall success within five attempts, on either part one or part two. These improvements are small enough that they do not cause changes in the significance relationships among conditions.

Looking only at no-storage participants, however, we do see another difference. As reported in Section 4.5, during part-one recall significantly more participants in pw-pronounce entered their secret correctly on the first try than in pp-large-3word. This is still true with the correction, and, in addition, pw-pronounce also performed better than pp-small (HC FET, p=.018). Part-two recall continues to have no significant difference (χ^2_{10}=7.252, p=0.701).

5.3 Off-by-One Errors

It is possible to construct a passphrase dictionary in which no word is within one edit of another. With such a dictionary, users who enter a word that is within one edit of the correct word in their passphrase can be authenticated successfully, with no loss of security. We did not attempt to create such a dictionary, but we did measure how many of our passphrase participants submitted entries with each word within one edit of the correct entry, as shown in Table 9. This correction is case-insensitive.

Applying this correction narrows the gap between passwords and passphrases. For first-attempt success during day-one recall, we still see an omnibus significant difference among conditions (χ^2_{10} =18.463, p=.048), but the pairwise differences showing greater accuracy for passwords than passphrases (see Section 4.5) disappear. Looking at only no-storage participants, however, pw-pronounce remains more successful than pp-large-3word (HC FET, p=.033).

Recall attempts on the second day continue to show no significant variation among conditions, including when we examine only no-storage participants.

5.4 Closest Dictionary Word Correction

Our security analysis assumes the attacker knows the dictionaries used to generate passphrases, and would therefore never guess a non-dictionary word. As a result, if a passphrase participant enters a word not included in the dictionary for his or her condition, we can replace the entered word with the closest (by edit distance) dictionary word, with no loss of security. Ties are arbitrarily but consistently broken using word order within the dictionary.

This correction can be applied only in our passphrase conditions and is case-insensitive; we did not implement it for our participants, but we examine how it would have affected their passphrase entries.

Results of our analysis are shown in Table 9. We find that this mechanism, like off-by-one correction, helps passphrase users some-

	All participants				No-storage			
	Part One		Part Two		Part One		Part Two	
Condition (correction)	On first try	In five tries	On first try	In five tries	On first try	In five tries	On first try	In five tries
pw-length5	95	99	83	99	93	98	72	98
pw-length6	90	99	82	99	90	95	60	100
pw-pronounce	95	99	83	99	94	98	75	98
– Ignore space+case	96	-	84	-	96	-	-	-
pp-small	81	95	79	95	65	85	62	92
– Ignore space+case	83	-	85	96	-	-	65	-
– Closest dict. word	85	-	82	96	69	-	65	-
– Edit distance	88	-	83	96	73	-	65	-
pp-medium	85	96	77	100	71	97	62	100
– Ignore space+case	88	-	80	-	76	-	-	-
– Closest dict. word	87	-	80	-	74	-	-	-
– Edit distance	87	-	81	-	74	-	-	-
pp-large	81	96	75	99	72	96	68	100
– Ignore space+case	87	-	81	-	80	-	-	-
– Closest dict. word	86	-	78	-	80	-	-	-
– Edit distance	91	97	79	-	88	-	-	-
pp-nouns	89	97	79	97	84	95	70	98
– Ignore space+case	90	-	82	-	-	-	75	-
– Closest dict. word	91	99	84	98	88	96	75	-
– Edit distance	91	98	84	98	88	96	75	-
pp-nouns-instr	88	97	74	98	80	93	61	97
– Ignore space+case	89	-	81	99	-	-	73	99
– Closest dict. word	91	-	77	-	83	-	66	-
– Edit distance	91	-	77	-	83	-	67	-
pp-sentence	89	98	78	98	85	100	70	93
– Ignore space+case	93	-	83	99	93	-	78	96
– Closest dict. word	90	99	82	99	-	-	74	96
– Edit distance	91	99	82	99	-	-	74	96
pp-large-3word	84	93	83	99	64	82	79	96
– Ignore space+case	87	94	89	-	68	86	82	-
– Closest dict. word	85	95	89	-	68	89	82	-
– Edit distance	-	94	89	-	-	86	82	-
pp-med-unorder	80	92	78	98	72	80	80	96
– Ignore space+case	81	-	81	99	-	-	-	-
– Closest dict. word	83	-	83	99	76	-	88	-
– Edit distance	85	93	85	99	80	84	88	-

Table 9: The percentage of participants in each condition who successfully recalled their secret in one and in five attempts in the first and second parts of the study. This includes participants who requested to have their secret emailed to them. The first row for each condition shows uncorrected data. Subsequent rows show the impact of correction for cases where correction would have allowed more users to log in; in many cases, correction did not help.

what but does not outperform uncorrected passwords. As with off-by-one correction, the only change we see in statistical relationships among conditions is for recall on day one; there remains an omnibus difference among conditions (χ^2_{10}=23.808, p=.008), but there are no longer any pairwise differences. For no-storage participants, likewise, we see omnibus significance in part one recall (χ^2_{10}=21.517, p=.018), but no pairwise significance.

5.5 Qualitative Error Analysis

We examined passphrase participants' recall errors in additional depth by manually categorizing error types. In this section, we focus on participants' first recall attempt during part two of the study.

Across all passphrase conditions, 214 participants (18.2% of those who completed day two) made errors during this attempt. Of these, we categorize 29.4% (63) as completely wrong — these entries had no apparent similarity to the participant's assigned passphrase. Most of these entries appear to be standard user-selected passwords, likely associated with another account belonging to the participant.

We next classify the errors made by the 151 passphrase participants whose entries did relate to their assigned passphrases (*related errors*). We believe that more than half of these errors could be mitigated relatively simply, with a passphrase scheme designed to tolerate common human errors.

In Section 5.2, we discuss ignoring spacing and capitalization errors. Among related errors, 25.2% (38) involved mistakes only in spacing, and another 7.9% (12) only in capitalization. Another 15.2% (23) of related errors appear to be simple typos; 21 of these could be corrected via the approach outlined in Section 5.3.

An additional 9.3% (14) of related errors involved the correct words appearing in the wrong order. Tolerating order variation reduces entropy, but our pp-med-unorder condition results suggest this entropy can be replaced by using a larger dictionary without sacrificing usability, at least up to a point. Another approach is to allow a user to succeed if he or she enters all but one of the words in the passphrase (e.g., [25]); as with variation in order, the lost entropy could be made up by using a larger dictionary. This technique, however, would have mitigated only 2.6% (4) of related errors.

Other errors made by our participants would require more sophisticated detection techniques. We identified 10 errors containing synonym substitutions. These synonyms, however, were almost all only loosely connected concepts — *war* to *army*, *position* to *proximity*, *friend* to *family*, *political* to *president* — suggesting that building a sufficiently synonym-tolerant dictionary without sacrificing security would be difficult More common than synonyms were rough sound-alikes, typically words with the same first letter and at least one similar vowel sound. Examples include *assort* for *according*, *over* for *officer*, *study* for *story*, and *meeting* for *morning*. We considered 17.2% of related errors (26) to include at least one error of this type. Although it is easy to understand how users make these errors, it is difficult to imagine a dictionary that could account for them without dramatically sacrificing security.

We hypothesized that participants would commonly confuse parts of speech — for example, typing *walk* instead of *walked* — or would substitute synonyms. While we did see occurrences of these errors, they were very limited.

Separating storage and no-storage participants reveals interesting trends. Among no-storage participants, 75.0% (69) of day-two, first-attempt recall errors were related; among storage participants, only 67.2% (82). This suggests that many storage participants either did not attempt to reference their stored passphrase, or else referenced the wrong stored secret. Among related errors, storage participants were more likely than non-storage participants to make spacing, capitalization, and typing errors, as might be expected when copying the passphrase from storage to the entry screen. While no-storage participants also made many of these simple errors, they were more likely than storage participants to make order, part of speech, and synonym errors; this is consistent with trying to recall the passphrase from memory. The frequency of sound-alike errors was similar for both storage and no-storage participants.

6. DISCUSSION

We compared the usability of eight types of system-assigned passphrases and three types of system-assigned passwords using a number of metrics, including memorability, time to authenticate, rate of user errors, tendency of users to store their secrets, and user sentiment. In this section we discuss ecological validity, and sum-

marize our high-level results about passphrases and some surprising findings about pronounceable passwords.

6.1 Ecological Validity

Several factors may affect the ecological validity and generalizability of our results. First, passphrases are unfamiliar to most users, whose behaviors and reactions might change given more experience. We assigned one, system-selected passphrase; we would expect different behavior from users with self-selected passphrases, and from those who keep track of multiple passphrases. Our memorability results are limited because so many participants stored their secrets. In addition, as mentioned in Section 3.1, our participants are also younger and more educated than the population at large.

Ecological validity in many password studies is limited by the fact that participants are aware they are using passwords for a study, rather than for accounts they value or expect to use long-term. We attempt to mitigate this by comparing only conditions that should be affected equally by this issue. In addition, we use a role-playing scenario, which our prior work suggests can motivate users to take their passwords more seriously than a survey scenario [32].

6.2 Summary of Results

System-assigned secrets. The use of system-assigned secrets eliminates the problem of users selecting low-entropy secrets, as well as the problem of users selecting a secret that they use for another account. We found that, in general, our system-assigned passwords and passphrases were not well-liked by users, and that the vast majority of users opted to store them. These results are consistent with the results of a previous study in which we found that 60% of participants stored their system-assigned 4-digit PINs [29]. In contrast, a study with similar methodology found that user-selected passwords were stored between 17% and 50% of the time, depending on condition [32]. Despite their unpopularity with users, system-assigned secrets may serve a role in situations where high entropy is a priority, and secure password storage poses minimal security risk and user inconvenience.

Dictionary choice. We experimented with passphrases composed of words drawn from a variety of dictionaries. All of the dictionaries we used were generated from the most frequently used words in COCA. We found that whether we used a dictionary of the top 181 words, top 401 words, or top 1,024 words made little difference for the metrics we studied. Using the top 181 nouns, or a sentence-like combination of the top 181 nouns, verbs, and adjectives also made little difference. This suggests that we may be able to create high-entropy passphrases while selecting dictionaries that meet certain properties, for example, dictionaries of words that are all at least three edits apart, which would allow the use of error correction to improve usability without sacrificing security.

Passphrase length. We found few differences between 3-word and 4-word passphrases. 3-word passphrases were shorter than 4-word passphrases drawn from the same dictionary, and therefore faster to type and resulted in fewer typing errors. But although 3-word passphrases were perceived as significantly less difficult to learn than several other conditions, the number of attempts needed to authenticate did not vary significantly. In addition, 3-word and 4-word passphrases with equivalent entropy are approximately the same length and result in similar typing speeds and error rates. For the conditions we studied, the number of characters in a passphrase appears to affect usability more than does the number of words.

Memory aids. We hypothesized that passphrases would be easier to remember if they were sentence-like, and that passphrases composed of nouns would be easier to visualize than passphrases composed of random words. However, we found that pp-sentence and pp-nouns resulted in slightly longer passphrases than pp-small (which contained short words such as: the, be, and, a, to), but otherwise performed similarly. We had also predicted that pp-nouns-instr would perform better than pp-nouns because the instructions would help people visualize and remember their passphrases. However, we found only small, statistically insignificant, differences between these two conditions. Different instructions, such as guiding users to visualize their passphrase or construct a scene or story using words from their passphrase, could prove more effective.

Word order. Requiring users to enter the words in a passphrase in a prescribed order increases the entropy of the passphrase. We explored whether this entropy increase came at the expense of usability. The pp-med-unorder condition was the same as the pp-medium condition, except it did not impose order requirements. Contrary to expectations, we did not find any significant differences between these conditions, nor between the pp-med-unorder and pp-small conditions, which used different dictionaries to maintain equivalent entropy. However, we found that participants did reorder their passphrases. 8.1% of participants in the pp-med-unorder condition took advantage of the ability to reorder their passphrases when entering them in the second part of the study (33.3% if we consider only no-storage participants). In passphrase conditions that did not permit reordering, we found that 9.3% of passphrase entry errors were due to entering words in the wrong order. Thus, it appears that relaxing the order requirement may provide small usability gains, but these gains were not significant in our study.

Error correction. Our analysis of the errors users made when entering their passphrases suggests that usability could be improved by selecting dictionaries that allow automatic correction of entry errors while maintaining a desired entropy. Even with the dictionaries we used, capitalization errors could be corrected without loss of entropy, because no two words differed only in capitalization. For the ordered passphrase conditions, missing spaces could also be corrected without loss of entropy. And, if every word in a dictionary had an edit distance of at least three from every other word in the dictionary, then it would be possible to correct many common typos as well as some errors where users misremember a word in their passphrase as another word that sounds similar.

Pronounceable passwords. We designed our experiment to compare system-assigned passphrases and passwords. Based on the negative sentiment and high storage rate associated with system-assigned passwords in a previous study [29], we were initially concerned that random-character system-assigned passwords might not provide a fair comparison. We looked for algorithms to generate relatively short, but high-entropy, system-assigned passwords that had characteristics that might make them more memorable. We found repeated mention in the literature of Gasser's algorithm for generating pronounceable random passwords [19]. However, members of our research group found passwords produced by this algorithm neither easy to pronounce nor easier to remember.

We were hence surprised to discover pw-pronounce performed very well — significantly better than some other conditions — in accuracy and entry-speed during part-one recall. One advantage of the pw-pronounce condition seems to be that the passwords in this condition include character combinations that, even if marginally pronounceable, are all lowercase and relatively easy to type.

7. ACKNOWLEDGMENTS

We thank Ashwini Rao for her feedback on study design. This research was supported in part by NSF grants DGE-0903659, CNS-1116776, and CCF-0424422; by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office; and by a gift from Microsoft Research.

8. REFERENCES

[1] A. Adams, M. A. Sasse, and P. Lunt. Making passwords secure and usable. In *Proc. HCI*, 1997.

[2] E. Adar. Why I hate Mechanical Turk research (and workshops). In *Proc. CHI Workshop on Crowdsourcing and Human Computation*, 2011.

[3] G. V. Bard. Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric. In *Proc. ACSW*, pages 117–124, 2007.

[4] A. J. Berinsky, G. A. Huber, and G. S. Len. Using Mechanical Turk as a subject recruitment tool for experimental research. *Political Analysis*, 2011.

[5] M. Bishop and D. V. Klein. Improving system security via proactive password checking. *Computers & Security*, 14(3):233–249, 1995.

[6] J. Bonneau. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *Proc. IEEE Symposium on Security and Privacy*, 2012.

[7] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of Web authentication schemes. In *Proc. IEEE Symposium on Security and Privacy*, 2012.

[8] J. Bonneau and E. Shutova. Linguistic properties of multi-word passphrases. In *Proc. USEC*, 2012.

[9] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Persp. Psych. Sci.*, 6(1):3–5, 2011.

[10] W. E. Burr, D. F. Dodson, and W. T. Polk. Electronic authentication guideline. Technical report, NIST, 2006.

[11] C. Castelluccia, M. Durmuth, and D. Perito. Adaptive password-strength meters from Markov models. In *Proc. NDSS*, 2012.

[12] D. Craddock. Hey! My friend's account was hacked! `http://windowsteamblog.com/windows_live/b/windowslive/archive/2011/07/14/hey-my-friend-s-account-was-hacked.aspx`, 2011.

[13] H. Crawford and J. Aycock. Kwyjibo: automatic domain name generation. *Softw: Pract. Exper.*, 38(14):1561–1567, 2008.

[14] M. Davies. The corpus of contemporary American English: 425 million words, 1990–present. Available online at `http://corpus.byu.edu/coca/`, 2008.

[15] M. Dell'Amico, P. Michiardi, and Y. Roudier. Password strength: An empirical analysis. In *Proc. INFOCOM*, 2010.

[16] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system? Screening Mechanical Turk workers. In *Proc. ACM CHI*, 2010.

[17] D. Florêncio and C. Herley. A large-scale study of web password habits. In *Proc. WWW*, 2007.

[18] A. Forget, S. Chiasson, P. C. van Oorschot, and R. Biddle. Improving text passwords through persuasion. In *Proc. SOUPS*, 2008.

[19] M. Gasser. A random word generator for pronouncable passwords. Technical Report ESD-TR-75-97, The MITRE Corporation, 1975.

[20] C. Herley and P. Van Oorschot. A research agenda acknowledging the persistence of passwords. *IEEE Security and Privacy*, 10(1):28–36, 2012.

[21] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 2010.

[22] InCommon Federation. Identity assurance profiles bronze and silver v1.1, 2011.

[23] P. Inglesant and M. A. Sasse. The true cost of unusable password policies: password use in the wild. In *Proc. ACM CHI*, 2010.

[24] P. G. Ipeirotis. Demographics of Mechanical Turk. Technical Report CeDER-10-01, New York University, 2010.

[25] M. Jakobsson and R. Akavipat. Rethinking passwords to adapt to constrained keyboards. *Proc. IEEE MoST*, 2012.

[26] S. Jeyaraman and U. Topkara. Have the cake and eat it too—Infusing usability into text-password based authentication systems. In *Proc. ACSAC*, 2005.

[27] M. Keith, B. Shao, and P. Steinbart. A behavioral analysis of passphrase design and effectiveness. *Journal of the Association for Information Systems*, 10(2):63–89, 2009.

[28] M. Keith, B. Shao, and P. J. Steinbart. The usability of passphrases for authentication: An empirical field study. *Int. J. Human-Comp. Studies*, 65(1):17–28, 2007.

[29] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. The impact of length and mathematical operators on the usability and security of system-assigned one-time PINs, 2012. Under review.

[30] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proc. IEEE Symp. Security & Privacy*, 2012.

[31] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proc. ACM CHI*, 2008.

[32] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: measuring the effect of password-composition policies. In *Proc. ACM CHI*, 2011.

[33] C. Kuo, S. Romanosky, and L. F. Cranor. Human selection of mnemonic phrase-based passwords. In *Proc. SOUPS*, 2006.

[34] S. A. Kurzban. Easily remembered passphrases: a better approach. *SIGSAC Rev.*, 3(2-4):10–21, Sept. 1985.

[35] K.-W. Lee and H.-T. Ewe. Passphrase with semantic noises and a proof on its higher information rate. In *Proc. CISW*, 2007.

[36] M. Leonhard and V. Venkatakrishnan. A new attack on random pronounceable password generators. In *Proc. IEEE EIT*, 2007.

[37] M. D. Leonhard and V. N. Venkatakrishnan. A comparative study of three random password generators. In *Proc. IEEE EIT*, 2007.

[38] K. Matsuura. Echo back in implementation of passphrase authentication. 2001.

[39] A. Mehler and S. Skiena. Improving usability through password-corrective hashing. In *Proc. SPIRE*, 2006.

[40] R. Munroe. xkcd: Password strength. `https://www.xkcd.com/936/`, 2012.

[41] NIST. Federal information processing standards publication 181: Automated password generator (APG). Technical report, 1993.

[42] S. N. Porter. A password extension for improved human factors. *Computers and Security*, 1(1), 1982.

[43] R. W. Proctor, M.-C. Lien, K.-P. L. Vu, E. E. Schultz, and G. Salvendy. Improving computer security for authentication of users: Influence of proactive password restrictions.

Behavior Res. Methods, Instruments, & Computers, 34(2):163–169, 2002.

[44] A. G. Reinhold. Diceware. http://world.std.com/~reinhold/diceware.html, 1995–2011.

[45] S. Riley. Password security: What users know and what they actually do. *Usability News*, 8(1), Feb. 2006.

[46] S. Schechter, C. Herley, and M. Mitzenmacher. Popularity is everything: a new approach to protecting passwords from statistical-guessing attacks. In *Proc. HotSec*, 2010.

[47] B. Schneier. Schneier on security blog. http://www.schneier.com/blog/archives/2005/06/write_down_your.html, 2005.

[48] S. Schoen, M. Hofmann, and R. Reynolds. Defending privacy at the U.S. border: A guide for travelers carrying digital devices. Electronic Frontier Foundation, 2011.

[49] A. Schumacher. Security @ CU—Making secure passwords, 2011.

[50] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423, 1949.

[51] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Proc. SOUPS*, 2010.

[52] A. Sotirakopoulos, I. Muslukov, K. Beznosov, C. Herley, and S. Egelman. Motivating users to choose better passwords through peer pressure. *SOUPS Poster*, 2011.

[53] Y. Spector and J. Ginzberg. Pass-sentence—a new approach to computer code. *Comput. Secur.*, 13(2):145–160, Apr. 1994.

[54] J. M. Stanton, K. R. Stam, P. Mastrangelo, and J. Jolton. Analysis of end user security behaviors. *Comp. & Security*, 24(2):124–133, 2005.

[55] M. Toomim, T. Kriplean, C. Pörtner, and J. Landay. Utility of human-computer interactions: toward a science of preference measurement. In *Proc. ACM CHI*, 2011.

[56] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How does your password measure up? The effect of strength meters on password creation. In *Proc. USENIX Security*, 2012. To appear.

[57] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proc. CCS*, 2010.

[58] S. Z. Wilson. The protect IU blog—xkcd agrees: Use a passphrase, 2011.

[59] J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Security and Privacy*, 2(5), Sept. 2004.

[60] Y. Zhang, F. Monrose, and M. K. Reiter. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Proc. CCS*, 2010.

[61] M. Zviran and W. J. Haga. A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal*, 36(3):227–237, 1993.

[62] M. Zviran and W. J. Haga. Password security: an empirical study. *J. Mgt. Info. Sys.*, 15(4), 1999.

APPENDIX

A. XKCD COMIC

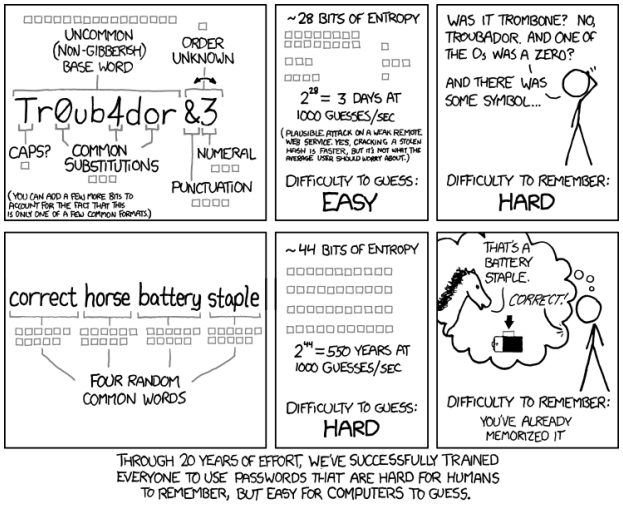

Figure 2: This xkcd comic suggests that users can recall passphrases composed of random words better than lower-strength passwords that meet complexity requirements [40].

B. DAY ONE SURVEY

This appendix includes the survey questions shown to participants during the first part of the study. All questions were required.

Learning my password was annoying.
( ) Strongly agree
( ) Agree
( ) Neutral
( ) Disagree
( ) Strongly disagree

Learning my password was difficult.
( ) Strongly agree
( ) Agree
( ) Neutral
( ) Disagree
( ) Strongly disagree

Learning my password was fun.
( ) Strongly agree
( ) Agree
( ) Neutral
( ) Disagree
( ) Strongly disagree

Describe anything you did to help yourself remember your password.

Do you have a password or set of passwords you reuse in different places?
( ) Yes
( ) No
( ) I prefer not to answer

Do you have a password that you use for different accounts with a slight modification for each account?
( ) Yes
( ) No
( ) I prefer not to answer
Do you have an email password?
( ) Yes
( ) No

The questions on this page pertain to your real email password.

What is the domain for your primary email account (e.g. hotmail.com, gmail.com, cmu.edu)?

Thinking about the real password you use for your primary email account, how many of the following does it contain? Write "0" if there are none.
Uppercase letters:
Lowercase letters:
Numbers:
Symbols:

Approximately how long ago did you last change your real email password?
( ) Within the past month
( ) Within the past six months
( ) Within the past year
( ) More than a year ago
( ) More than 5 years ago
( ) Never
( ) I'm not sure
( ) I prefer not to answer

Does your main email provider require you to change your password periodically?
( ) Yes
( ) No
( ) I'm not sure
( ) I prefer not to answer

If my main email account assigned me a password like the one I used in this study, it would make my email account more secure.
( ) Strongly agree
( ) Agree
( ) Neutral
( ) Disagree
( ) Strongly disagree

I would be annoyed if my main email account assigned me a password like the one I used in this study.
( ) Strongly agree
( ) Agree
( ) Neutral
( ) Disagree
( ) Strongly disagree

If my main email account assigned me a password like the one I used in this study, it would be easier.
( ) Strongly agree
( ) Agree
( ) Neutral
( ) Disagree
( ) Strongly disagree

Are you willing to return and try to recall your password again in a few days?
( ) Yes
( ) No
( ) I prefer not to answer

If you have any additional feedback about passwords or this survey, please enter your comments here.

What is your gender?
( ) Female
( ) Male
( ) I prefer not to answer

How old are you?

Which of the following best describes your highest achieved education level?
( ) Some High School
( ) High School Graduate
( ) Some college, no degree
( ) Associates degree
( ) Bachelors degree
( ) Graduate degree (Masters, Doctorate, etc.)
( ) Other

Are you majoring in or do you have a degree or job in computer science, computer engineering, information technology, or a related field?
( ) Yes
( ) No
( ) I prefer not to answer

Are you majoring in or do you have a degree or job in art, architecture, design, photography, or a related field?
( ) Yes
( ) No
( ) I prefer not to answer

Are you majoring in or do you have a degree or job in math, physics, or engineering, or a related field?
( ) Yes
( ) No
( ) I prefer not to answer

What is your total household income?
( ) Less than $10,000
( ) $10,000 to $19,999
( ) $20,000 to $29,999
( ) $30,000 to $39,999
( ) $40,000 to $49,999
( ) $50,000 to $59,999
( ) $60,000 to $69,999
( ) $70,000 to $79,999
( ) $80,000 to $89,999
( ) $90,000 to $99,999
( ) $100,000 to $149,999
( ) $150,000 or more
( ) Prefer not to answer

Thank You!

C. DAY TWO SURVEY

This appendix includes the survey questions shown to participants during the second part of the study. All questions were required.

Thank you for participating in this Carnegie Mellon University study. Please answer the following questions honestly. There are no right or wrong answers and everyone who finishes this task completely will receive his or her bonus payment.

How did you just enter your password for this study (please be honest – you get paid regardless, and this will help our research)?
( ) I typed it in from memory
( ) It was stored in my browser
( ) I cut and pasted it from a text file
( ) I looked it up in the place I had recorded it earlier and then I typed it in
( ) I use a password manager that filled it in for me
( ) I prefer not to answer
( ) Other:
( ) It was automatically filled in
( ) I forgot my password and followed the password reset link

Did you write down or store the password you created for this study (please be honest, you get paid regardless, this will help our research)?

[ ] No
[ ] Yes, on paper
[ ] Yes, electronically (stored in computer, phone, etc.)
[ ] Other
[ ] I prefer not to answer

If you wrote down or stored your password for this study, how is it protected (choose all that apply)?
[ ] I do not protect it
[ ] I stored it in an encrypted file
[ ] I hid it
[ ] I stored it on a computer or device protected with another password
[ ] I locked up the paper
[ ] I always keep the password with me
[ ] I wrote down a reminder instead of the actual password
[ ] I keep the paper in an office or room that only I use
[ ] I stored it on a computer or device that only I use
[ ] Other
[ ] I prefer not to answer
[ ] I did not write down my password

Please describe how you store your password for this study, including what software you use or where you wrote it down.

What would you have done differently in protecting and remembering your password if this password were used for an account you would use outside this study?

Did you imagine a scene related to the words or letters in your password to help you remember it?
( ) Yes
( ) No

If so, describe the scene that you imagined.

Did you think of a sentence or phrase based on the words or letters in your password to help you remember it?
( ) Yes
( ) No

If so, describe the sentence or phrase that you used.

Did you think of a story related to the words or letters in your password to help you remember it?
( ) Yes
( ) No

If so, describe the story that you used.

What, if anything, about your new password makes it easy for you to remember?

Do you have an email password?
( ) Yes
( ) No

The questions on this page pertain to your real email password.

When logging in with your real email password, do you refer to a written down or stored password?
( ) Yes
( ) No

Prior to this survey, have you ever written down or stored your real email password?
[ ] No
[ ] Yes, on paper
[ ] Yes, electronically (stored in computer, phone, etc.)
[ ] Other
[ ] I prefer not to answer

If you ever wrote down or stored your real email password, how was it protected (choose all that apply)?
[ ] I did not write down or store it
[ ] I did not protect it
[ ] I stored it in an encrypted file
[ ] I hid it
[ ] I stored it on a computer or device protected with another password
[ ] I locked up the paper
[ ] I always kept the password with me
[ ] I wrote down a reminder instead of the actual password
[ ] I kept the paper in an office or room that only I use
[ ] I stored it on a computer or device that only I use
[ ] Other
[ ] I prefer not to answer

To how many people have you given your real email password?
( ) 0
( ) 1
( ) 2-5
( ) 6-10
( ) More than 10

Consider the password you used for this study. If you were protecting/remembering a password for a real email account you would use outside the study, what would you have done differently?
[ ] Nothing would have changed
[ ] I would have written it down on paper
[ ] I would not have written it down on paper
[ ] I would have stored it electronically
[ ] I would not have stored it electronically
[ ] I would still write it on paper, but would secure the paper better
[ ] I would have tried harder to remember it
[ ] Other

How often do you type in your real email password (we are interested in when you type it in, not when your browser enters it automatically)?
( ) Never
( ) Several times per day
( ) Once per day
( ) Several times per week
( ) Once per week
( ) A few times per month
( ) Once per month
( ) Less than once a month
( ) I prefer not to answer

Remembering the password I use for my real email account is difficult.
( ) Strongly agree
( ) Agree
( ) Neutral
( ) Disagree
( ) Strongly disagree
Remembering the password I used for this study was difficult.
( ) Strongly agree
( ) Agree
( ) Neutral
( ) Disagree
( ) Strongly disagree

If you have any additional feedback about passwords or this survey, please enter your comments here.

Thank you.

D. CONDITIONS

This appendix includes details about the password and passphrase construction for each condition introduced in Section 3.2 First we detail the characters used for random password construction, then we enumerate the terms found in each dictionary for conditions that used dictionaries. The ordering of the terms in each dictionary is by frequency as calculated by the Corpus of Contemporary American English [14]. The pronounceable condition uses the algorithm described in [19].

D.1 Random character passwords

The random character conditions used passwords created from the following set of characters:

```
abcdefghjkmnpqrstuvwxyz
ABCDEFGHJKLMNPQRSTUVWXYZ
23456789
@!$*#.-&_
```

This set specifically omits characters that could be confused with other characters, such as mistaking the number zero for the letter

'O' or the number one for an uppercase 'I' or lowercase 'L' character.

D.2 181-word dictionary

Our 181 word dictionary, used in the pp-small condition, consists of the following words.

the be and of a in to have to it I that for you he with on do say this they at but we his from that not by she or as what go their can who get if would her all my make about know will as up one time there year so think when which them some me people take out into just see him your come could now than like other how then its our two more these want way look first also new because day more use no man find here thing give many well only those tell one very her even back any good woman through us life child there work down may after should call world over school still try in as last ask need too feel three when state never become between high really something most another much family own out leave put old while mean on keep student why let great same big group begin seem country help talk where turn problem every start hand might American show part about against place over

D.3 401-word dictionary

Our 401 word dictionary, used in pp-medium and pp-med-unorder conditions, consists of the following words.

the be and of a in to have to it I that for you he with on do say this they at but we his from that not by she or as what go their can who get if would her all my make about know will as up one time there year so think when which them some me people take out into just see him your come could now than like other how then its our two more these want way look first also new because day more use no man find here thing give many well only those tell one very her even back any good woman through us life child there work down may after should call world over school still try in as last ask need too feel three when state never become between high really something most another much family own out leave put old while mean on keep student why let great same big group begin seem country help talk where turn problem every start hand might American show part about against place over such again few case most week company where system each right program hear so question during work play government run small number off always move like night live Mr point believe hold today bring happen next without before large all million must home under water room write mother area national money story young fact month different lot right study book eye job word though business issue side kind four head far black long both little house yes after since long provide service around friend important father sit away until power hour game often yet line political end among ever stand bad lose however member pay law meet car city almost include continue set later community much name five once white least president learn real change team minute best several idea kid body information nothing ago right lead social understand whether back watch together follow around parent only stop face anything create public already speak others read level allow add office spend door health person art sure such war history party within grow result open change morning walk reason low win research girl guy early food before moment himself air teacher force offer enough both education across although remember foot second boy maybe toward able age off policy everything love process music including consider appear actually buy probably human wait serve

D.4 1024-word dictionary

Our 1024 word dictionary, used in pp-large and pp-large-3word conditions, consists of the following words.

the be and of a in to have to it I that for you he with on do say this they at but we his from that not by she or as what go their can who get if would her all my make about know will as up one time there year so think when which them some me people take out into just see him your come could now than like other how then its our two more these want way look first also new because day more use no man find here thing give many well only those tell one very her even back any good woman through us life child there work down may after should call world over school still try in as last ask need too feel three when state never become between high really something most another much family own out leave put old while mean on keep student why let great same big group begin seem country help talk where turn problem every start hand might American show part about against place over such again few case most week company where system each right program hear so question during work play government run small number off always move like night live Mr point believe hold today bring happen next without before large all million must home under water room write mother area national money story young fact month different lot right study book eye job word though business issue side kind four head far black long both little house yes after since long provide service around friend important father sit away until power hour game often yet line political end among ever stand bad lose however member pay law meet car city almost include continue set later community much name five once white least president learn real change team minute best several idea kid body information nothing ago right lead social understand whether back watch together follow around parent only stop face anything create public already speak others read level allow add office spend door health person art sure such war history party within grow result open change morning walk reason low win research girl guy early food before moment himself air teacher force offer enough both education across although remember foot second boy maybe toward able age off policy everything love process music including consider appear actually buy probably human wait serve market die send expect home sense build stay fall oh nation plan cut college interest death course someone experience behind reach local kill six remain effect use yeah suggest class control raise care perhaps little late hard field else pass former sell major sometimes require along development themselves report role better economic effort up decide rate strong possible heart drug show leader light voice wife whole police mind finally pull return free military price report less according decision explain son hope even develop view relationship carry town road drive arm true federal break better difference

thank receive value international building action full model join season society because tax director early position player agree especially record pick wear paper special space ground form support event official whose matter everyone center couple site end project hit base activity star table need court produce eat American teach oil half situation easy cost industry figure face street image itself phone either data cover quite picture clear practice piece land recent describe product doctor wall patient worker news test movie certain north love personal open support simply third technology catch step baby computer type attention draw film Republican tree source red nearly organization choose cause hair look point century evidence window difficult listen soon culture billion chance brother energy period course summer less realize hundred available plant likely opportunity term short letter condition choice place single rule daughter administration south husband Congress floor campaign material population well call economy medical hospital church close thousand risk current fire future wrong involve defense anyone increase security bank myself certainly west sport board seek per subject officer private rest behavior deal performance fight throw top quickly past goal second bed order author fill represent focus foreign drop plan blood upon agency push nature color no recently store reduce sound note fine before near movement page enter share than common poor other natural race concern series significant similar hot language each usually response dead rise animal factor decade article shoot east save seven artist away scene stock career despite central eight thus treatment beyond happy exactly protect approach lie size dog fund serious occur media ready sign thought list individual simple quality pressure accept answer hard resource identify left meeting determine prepare disease whatever success argue cup particularly amount ability staff recognize indicate character growth loss degree wonder attack herself region television box TV training pretty trade deal election everybody physical lay general feeling standard bill message fail outside arrive analysis benefit name sex forward lawyer present section environmental glass answer skill sister PM professor operation financial crime stage ok compare authority miss design sort one act ten knowledge gun station blue state strategy little clearly discuss indeed force truth song example democratic check environment leg dark public various rather laugh guess executive set study prove hang entire rock design enough forget since claim note remove manager help close sound enjoy network legal religious cold form final main science green memory card above seat cell establish nice trial expert that spring firm Democrat radio visit management care avoid imagine tonight huge ball no close finish yourself talk theory impact respond statement maintain charge popular traditional onto reveal direction weapon employee cultural contain peace head control base pain apply play measure wide shake fly interview manage chair fish particular camera structure politics perform bit weight suddenly discover candidate top production treat trip evening affect inside conference unit best style adult worry range mention rather far deep past edge individual specific writer trouble necessary throughout challenge fear shoulder institution middle sea dream bar beautiful property instead improve stuff detail method sign somebody magazine hotel soldier reflect heavy sexual cause bag heat fall marriage tough sing surface purpose exist pattern whom skin agent owner machine

D.5 Noun dictionary

Our 181 noun dictionary, used in pp-nouns, pp-nouns-instr, and pp-sentence conditions, consists of the following words.

time year people way day man thing woman life child world school state family student group country problem hand part place case week company system program question work government number night Mr point home water room mother area money story fact month lot right study book eye job word business issue side kind head house service friend father power hour game line end member law car city community name president team minute idea kid body information back parent face others level office door health person art war history party result change morning reason research girl guy food moment air teacher force education foot boy age policy process music market sense nation plan college interest death experience effect use class control care field development role effort rate heart drug show leader light voice wife police mind price report decision son view relationship town road arm difference value building action model season society tax director position player record paper space ground form event official matter center couple site project activity star table need court American oil situation cost industry figure street image phone

D.6 Sentence-like dictionaries

The sentence like condition is constructed from a noun followed by a transitive verb, then an adjective and finally another noun. Due to this construction a dictionary is required for each part of speech type. For nouns, the same 181 words pp-nouns dictionary is used. For verbs, the following 181 words are used.

has does says goes cans gets makes knows wills thinks takes sees wants looks uses finds gives tells works calls tries asks needs feels becomes leaves puts means keeps lets begins helps talks turns starts shows hears plays runs moves likes lives believes holds brings writes provides sits stands loses pays meets includes continues sets learns changes leads understands watches follows stops creates speaks reads allows adds spends grows opens walks wins offers remembers loves considers buys waits serves dies sends expects builds stays falls cuts reaches kills suggests raises passes sells requires reports decides pulls returns explains hopes develops carries drives breaks thanks receives joins agrees picks wears supports ends hits bases produces eats teaches faces covers describes catches draws chooses causes points realizes places closes involves increases seeks deals fights throws fills represents focuses drops plans pushes reduces notes enters shares rises shoots save protects lies accepts identifies determines prepares argues recognizes indicates wonders lays fails names presents answers compares misses acts states discusses forces checks laughs guesses studies proves hangs designs forgets claims removes sounds enjoys forms establishes

For adjectives, the following 181 words are used.

other new good high old great big American small large national young different black long little important political bad white real best right social only public sure low early able human local late hard major better economic strong possible whole free military true federal international full special easy clear recent certain personal open red difficult available likely short single medical current wrong private past foreign fine common poor natural significant similar hot dead central happy serious ready simple left physical general environmental financial blue democratic dark various entire close legal religious cold final main green nice huge popular traditional cultural wide particular top far deep individual specific necessary middle beautiful heavy sexual tough commercial total modern positive civil safe interesting rich western senior key professional successful southern fresh global critical concerned effective original basic powerful perfect involved nuclear British African very sorry normal Chinese front supposed Soviet future potential European independent Christian willing previous interested wild average quick light bright tiny additional present warm annual French responsible regular soft female afraid native broad wonderful growing Indian quiet aware complete active chief