Improving security and usability of passphrases with guided word choice

Nikola K. Blanchard, Clément Malaingre, Ted Selker

ABSTRACT

Passphrases have many uses, such as serving as seeds for passwords. User-created passphrases are easier to remember, but tend to be less secure than ones created from words randomly chosen in a dictionary. This paper develops a way of making more memorable, more secure passphrases. It investigates the security and usability of creating a passphrase by choosing from a randomly generated set of words presented as a two-dimensional array. A usability experiment shows that participants using this method achieved 97% to 99% of the maximal theoretical entropy and commited fewer than half as many memory mistakes as a control group with assigned passphrases. It also shows that their choices are affected by word familiarity and weakly by the word's position in the array. Prompting a person with random words from a large dictionary is an effective way of helping them make a more memorable high-entropy passphrase.

CCS CONCEPTS

• **Security and privacy → Authentication**; *Usability in security and privacy*; • **Human-centered computing → Empirical studies in HCI**;

KEYWORDS

Usable security; Mnemonic phrases; Passwords

ACM Reference Format:
Nikola K. Blanchard, Clément Malaingre, Ted Selker. 2018. Improving security and usability of passphrases with guided word choice. In *2018 Annual Computer Security Applications Conference (ACSAC '18), December 3–7, 2018, San Juan, PR, USA.* ACM, New York, NY, USA, Article 39, 10 pages. https://doi.org/10.1145/3274694.3274734

1 INTRODUCTION

It is hard to make high entropy (highly random) codes, and it is hard to remember them. Typical uses of passwords have suffered from serious usability and security problems [Cranor 2014, 2016]. Low-entropy selection methods, poor memorability, and rules that make passwords difficult to retrieve all reduce their utility. Biometric methods are promising, but they still suffer from many vulnerabilities, typically being hacked within six months of introduction [Cao and Jain 2016; Reddy et al. 2008; Ruiz-Albacete et al. 2008; Smith et al. 2015]. Biometric security approaches also have an increased risk of unmitigatable leaks about a particular user [Simoens et al. 2009] (as a retina is harder to change than a password). As well as the typical HCI research, the usability of systems is deeply effected by our access problems.

Systems have a critical need for security alternatives that can handle sharing and changing the access method (such as giving a database password to someone for the evening). Longer lists of words – or passphrases – have been suggested as a possible solution for improving security [Shay et al. 2014; Topkara et al. 2007]. Although they were introduced as early as 1982 [Porter 1982], dozens of years later passphrases still suffer from problems similar to the ones passwords have.

Passphrases have been shown to typically be made from insecure, linguistically easy-to-crack patterns [Kuo et al. 2006], like song lyrics or famous quotes. A significant number of passphrases created by the Amazon PayPhrase system were easily hackable, with 1.3 percent of accounts being vulnerable to a 20,000-word dictionary of terms used in popular culture [Bonneau 2012; Bonneau and Shutova 2012]. In another study [Yang et al. 2016], one passphrase method led 2.55% of users to choose the same sentence. Despite multiple protocols encouraging users to make personalized sentences, five out of six methods had many occurrences of different users ending up with the same passphrase. With the single method that had no two users choose the same password, they still observed some quotes and famous sentences, but few enough not to have collisions on a database of only 777 passwords. These studies show that letting people choose a passphrase with no constraints leads to low-entropy passphrases, even when giving them instructions to make personalized passphrases.

Recent work has debated the value of entropy as a measure of password strength [Ma et al. 2010; Rass and König 2018; Taha et al. 2013]. However, those concerns are not directly transposable to passphrases, as the main use-case for those isn't as a replacement for passwords (which would require too many passphrases), but as tools that can serve as sources of entropy for other password generation methods[Blanchard et al. 2018; Blocki et al. 2014]. Hence, the usefulness of passphrases depends on them having high entropy. To achieve this, they have to be not only longer but, more importantly, less predictable. As opposed to complex passwords which are hard to remember [Komanduri et al. 2011; Pilar et al. 2012; Yan et al. 2004], passphrases benefit directly from our natural abilities to remember sequences of words [Baddeley and Hitch 1974; Miller 1956]. High-entropy passphrases can then serve many purposes, such as seeding diverse passwords [Blanchard et al. 2018; Blocki et al. 2014] or creating other codes, while avoiding the pitfalls of password reuse [Lipa 2016; Segreti et al. 2017; Wash et al. 2016].

To avoid users choosing common passphrases, one could take inspiration from standard password practice and draw words uniformly from a dictionary. However, this has two drawbacks: first, the passphrases generated are not individualized and can suffer from low memorability. Second, to make sure that the user knows all the words generated, the dictionary from which the words are drawn must be of limited size.

ACSAC '18, December 3–7, 2018, San Juan, PR, USA

ACM ISBN 978-1-4503-6569-7/18/12…$15.00
https://doi.org/10.1145/3274694.3274734

This paper explores a method of guiding the user to choose their passphrase from an imposed set of random words. A usability experiment explores the factors affecting the word choice, the participants' ability to remember their passphrases, and the type of mistakes they make. Both for primary and secondary English speakers, this method leads to highly increased memorability of the passphrases created. Moreover, since the user can choose from several words, the dictionary need not be made of only words that are sure to be known to a user, allowing the use of much bigger dictionaries, leading to an increase not only in usability but also in security.

The authors would like to thank Amira Lakdhar for detecting and allowing us to fix an important error in the data treatment pipeline, and Irène Gannaz for her advice and help with the statistical analysis.

2 METHOD

An online usability experiment explored the impact of creating a passphrase by choosing words from an array.

2.1 Word choice

Presented guide words were drawn uniformly (each word having the same probability) from a dictionary crafted for this purpose. This dictionary is based on the first third of Peter Norvig's 300,000 most frequent n-grams [Norvig 2009]. As those 100,000 words still included words from other languages such as "unglaublichen" as well as some non-words like "unixcompile", only ones which were also in the SOWPODS (list of admissible words in English Scrabble tournament) were kept. This created a list of the 87,691 most frequent English words. Thanks to shared roots in words, there is evidence that most people would know a large majority of them [Brysbaert et al. 2016; Hartmann 1946]. As the participants chose only 6 words from the 20 or 100 word arrays, having a few unknown words in it shouldn't in any case change the process or outcome.

To get a variety of behaviours, we experimented with word arrays of different sizes. To give participants a real choice, the guide array to choose from was created to be several times the length of the passphrase created from it. It also needed to not be so long that people got overwhelmed, had to scroll through them, or took too much time making decisions. An array of size 100 was the biggest that could reliably fit on a computer screen, and 20 was the smallest that guaranteed enough possibilities for the user to choose from (38,760 in total, discounting order).

2.2 Protocol

The experiment was hosted on a privately hosted server and accessed remotely over the web. It ran on the Scala Play framework and a PostgreSQL database. Participants were shown the following pages, with instructions at the top of each:

(1) A welcome page that gave participants an overview of the activity and informed them of their rights.
(2) A question asking their age and another asking the primary language they used.
(3) A dynamically generated array of either 20 or 100 words (A/B testing). Participants were told to select 6 words, in the order of their choice, and input them in the 6 text-boxes at the bottom of the page. The words were presented in five columns of either 4 or 20 words that were not aligned with the input boxes to limit the potential bias of choosing one word per column.
A control experiment was run to create a baseline for remembering a 6-word passphrase. Instead of choosing their words from an array, a sequence of 6 randomly generated words was directly given, while informing them that it had been randomly created.
(4) A page that repeated the passphrase, then prompted participants to repeat it to themselves until they could remember it.
(5) A text-box was presented with instructions to type in the first two letters of each of their words.
(6) A page was presented, showing an array of words that had previously been presented to another participant. They were then told to try guessing what words the other participant had chosen.
(7) A page informing them which if any of their guessed words were correct, and telling them that they could try to guess more passphrases if they wanted, or could continue with the rest of the experiment.
(8) A page asking them to repeat all six words from their passphrase in the same order, or as many as they could think of if they didn't remember all of them. If some were missed, they were then presented with their original array of words as a clue and asked to find all six of them .
(9) A page thanking them for their participation and inviting them to encourage others to become participants.

The experiment collected of the following data for analysis:

- Any information entered in the text-boxes.
- All the words and arrays shown to the participants.
- Time spent on each page.
- List of (keystroke/timestamp) couples.

To make sure that no one would try the experiment multiple times to improve their performance, IP addresses associated with the participants were temporarily kept. A single occurrence of a second try by a participant was detected and was excluded from the database.

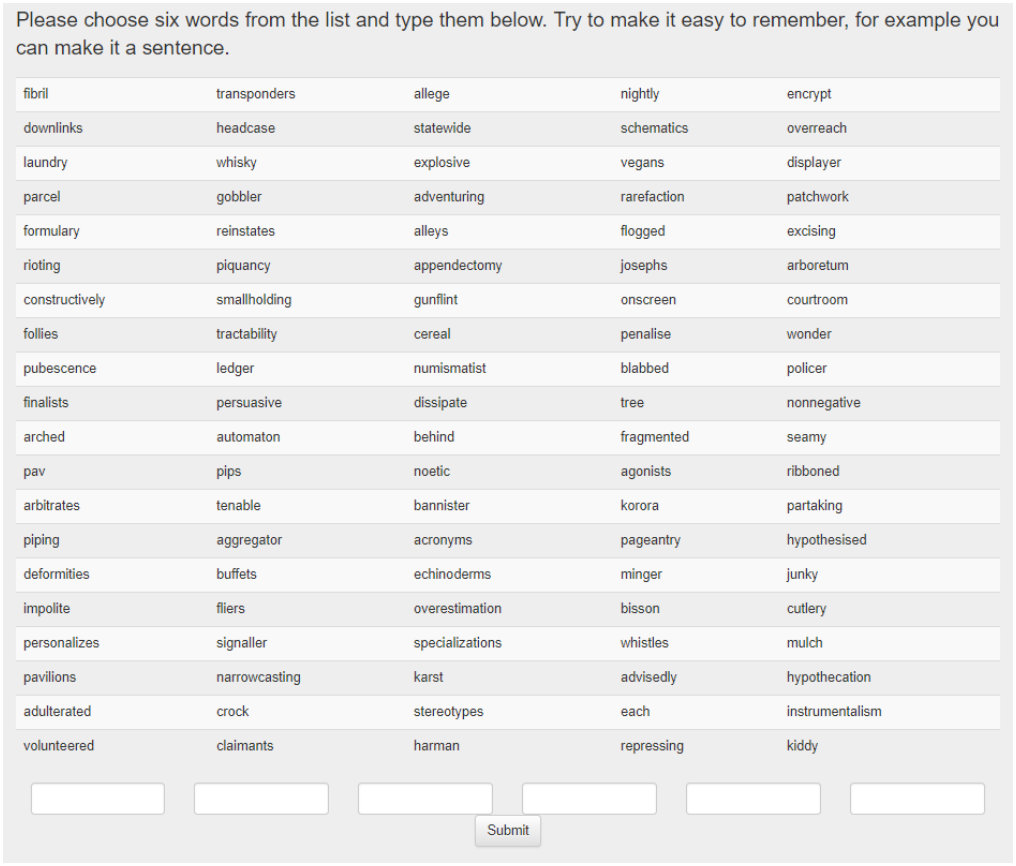
2.3 Design choices

Passphrases of length 6 were chosen as they provide the entropy required in previous work on password generation from passwords. Moreover, this is compatible with known bounds on memory and information processing ability [Miller 1956].

The guide array was purposely designed not to line up with the spaces for words in the passphrase below (as shown on Figure 1) to avoid confusion in step 3 and separate the guide words from the user-chosen passphrase. Participants were also told to try to make their passphrase memorable, for example, by creating a phrase, rhyme, or sentence from their selected words.

Step 5 was meant to help learn the passphrase and check whether it was memorized. Step 6 was then introduced as a distractor exercise introduced to interfere with their short term memory for passphrases. The idea was to eliminate short term memory of their initial passphrase by making them think of someone else's passphrase.

Figure 1: Screenshot of the word choosing interface



3 DEMOGRAPHIC INFORMATION

3.1 Participant selection

The principles of informed consent (including right to quit and right to privacy), not using people in protected classes, beneficence, justice and minimal deception were followed. All the participants were volunteers, and were informed of the length of experiment and that they could quit at any point. They were told that it was an opportunity to help them test their memory and for us to understand how people typed, and that their typing would be monitored. For privacy, minimal demographic data was collected, corresponding to aspects that would be relevant to analysing results.

3.2 Recruitment of volunteers

Volunteers were recruited through John Krantz's Psychological Research on the Net website[Krantz 1998] which promotes and indexes experiments of this kind. These volunteers were also encouraged to invite others to participate. Late volunteers were hence partially recruited through social networks. Recruited participants were not members of protected classes and were informed that they could leave at any point. They were not asked any identifying information and were informed not to use their input for security purposes as their data would be collected for analysis.

3.3 Statistics

Groups. A total of 125 people participated and were randomly assigned to three groups. Group "20" was shown an array of 20 words to choose from and was composed of 47 volunteers. Group "100" was shown an array of 100 words and had 52 volunteers. The control group, with 26 volunteers, had their words chosen for them and shown on the screen.

Age. The participants' ages showed a large variation, from 16 to 69, with a notable concentration around 24. The average was 31 years old, and the median 25.

Language. 51 participants wrote down English as their primary language. French (28) and Hebrew (14) were the next two most reported primary languages, followed by Arabic, Norwegian, Russian, and Romanian.

4 RESULTS

The results focus on how participants chose words and how different variables affected their choices, on their ability to remember the words chosen, and on how they guessed other people's words.

4.1 Word selection

Based on problems recognized in other studies[Blunch 1984; Lerman and Hogg 2014; Payne Stanley 1951; Yang et al. 2016; Yue et al. 2010], an original hypothesis was that word choice would be influenced by three behaviours:

- *Semantic*: participants might choose words that are more frequently used (and with which they were more familiar);
- *Syntactic*: participants might choose words that are compatible with others they chose, to create a sentence with a common structure.
- *Positional*: participants might choose words that are either among the first they read (on the top left corner), or closest to the input fields;

A possible outcome that would make the proposed method ineffective – because of low entropy – is that people could all choose the most familiar or frequent words from the array of n words shown to them. Similarly, strong syntactic tendencies could lower entropy by reducing the number of probable passphrases. On the other hand, positional effects tend to increase the entropy as they don't depend on human biases but on random position, making the distribution more uniform.

Results below showed that the second effect is much weaker than could be expected, with some choosing bias caused by semantic and positional effects.

4.1.1 Semantic effects. To analyse the semantic effects, we used a dictionary with words sorted by decreasing order of frequency in the n-words corpus. Frequent words are all within the first few thousand words in this dictionary, with rare words at the end.

The histograms in Figure 2 show the distribution of the words chosen depending on their frequency rank in the dictionary, for each of the two main groups. To make the figure more legible, ranks were aggregated in 30 buckets of 2923 words. Although we can observe a bias in favour of more frequent and familiar words, 23% of words chosen still come from the least frequent half of the dictionary; group "20" used rare words 26% of the time, group "100" used rare words 20% of the time.

The significant fraction of words chosen from the second half of the dictionary is not just due to some participants getting only rare words, as the following figures show. Figures 3 and 4 show the distribution of words chosen depending on their frequency relative to the frequencies of the words shown to the participant. This frequency, F(i), is equal to the number of times the i-th most frequent word in the array was chosen. Those histograms also include show the distinction between primary English speakers and others, inside each group. When given enough choice – in group "100" – non-primary speakers have a bigger tendency to choose frequent words, with only 15% choosing rarer words. A single participant in group "20" chose the 6 most frequent words in their array, which is more than the expectation of uniform random choice of words[1].

4.1.2 Syntactic effects. Using common sentence structures might help memorization. Some participants apparently tried to make use of this, one of them choosing "Freshman minions cinematically crumble lavender prints". However, most participants did not try to make a gramatical passphrase. Even when limiting the analysis to the four broadest grammatical categories (noun/verb/adjective/adverb), most sentence structures were unique, with 67 grammatical structures seen only once out of 99 passphrases. The grammatical structure in the example passphrase (noun-verb-noun-noun-verb-noun) is among those unique structures. 9 structures were seen twice and 3 were seen thrice. The only relatively common structure, which was present 8 times, corresponds to a sequence of 6 nouns.

Figure 5 shows how present each grammatical category was in each position in the passphrase. For example, adjectives are less present as a second word than as a first. There is some imprecision as words can fit in multiple categories (e.g. *scars* as a noun or a verb). Nouns seem overrepresented, but this is consistent with their frequency in the dictionary (≈ 60%). No correlations were found between successive word categories, which would be difficult in any case because of the multiple potential roles of each word and additional inquiry should be conducted on the subject.

[1]Choosing 6 words at random from an array of 20 gives any given outcome with probability 0.003%.

Figure 2: Frequencies of the words chosen by each group as a function of their rank in the dictionary, by buckets of 2923 words

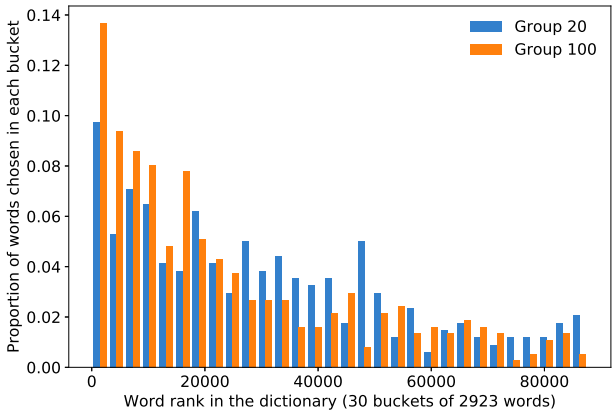


Figure 3: Relative word frequencies for group "20"

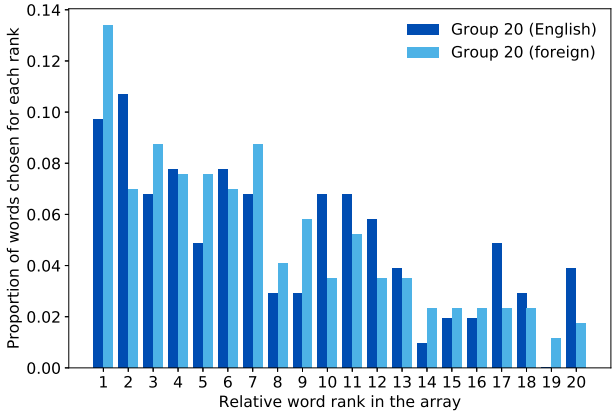


Figure 4: Relative word frequencies for group "100"

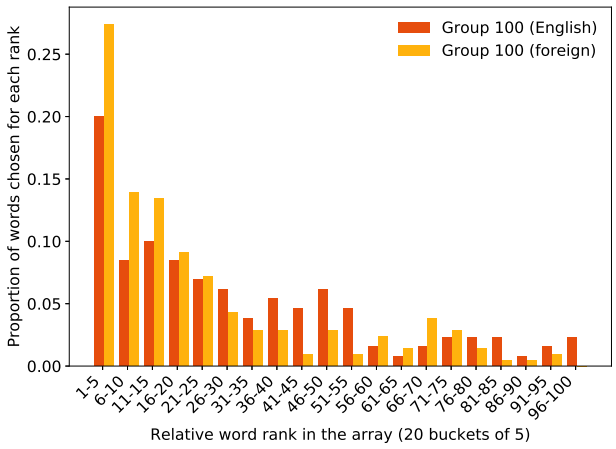
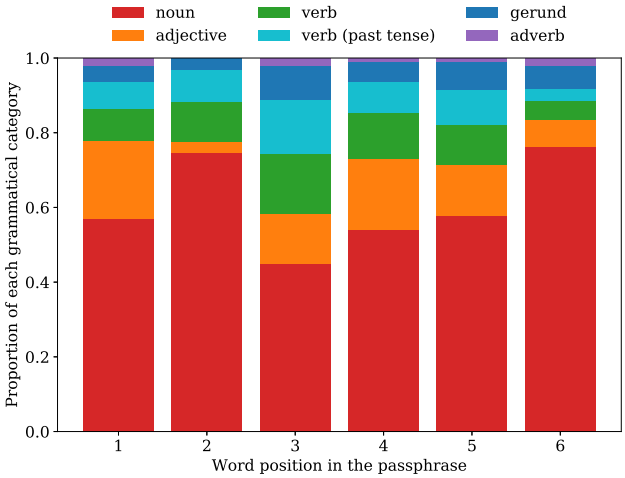


Figure 5: Repartition of grammatical categories by position in passphrase



In group "20", 12 people created passphrases that made some amount of semantic sense and followed English syntax, such as the example given. 13 passphrases could make some sense but had unusual or incorrect syntax, such as "furry grills minidesk newsdesk deletes internet". 22 appeared to be six randomly ordered words, such as "wastewater refundable sweatshops misspelling sellout ailment ". In group "100", only 6 people created passphrases that made some amount of sense and were syntactically correct. 15 made passphrases that could make some sense, and 30 had passphrases that seemed randomly ordered.

4.1.3 Positional effects. The heat maps in Figures 6 and 7 show how the position of a word in the array shown affected the probability that it would get chosen. The numbers correspond to the percentage of participants who chose the word in that cell, with a deeper red indicating a higher percentage. The numbers beside and below the heat maps represent the total number of words chosen by line and column.

Figure 6: Heatmap indicating the percentage of participants choosing the word in each cell for group "100"

5.6	5.6	10.1	2.3	11.3	31
1.1	3.4	1.1	3.4	10.1	17
6.8	5.6	1.1	1.1	1.1	14
6.8	5.6	9.0	9.0	4.5	31
1.1	6.8	3.4	6.8	4.5	20
9.0	11.3	6.8	5.6	5.6	34
6.8	2.3	1.1	9.0	4.5	21
9.0	6.8	10.1	3.4	5.6	31
10.1	7.9	10.1	3.4	5.6	33
6.8	9.0	10.1	3.4	3.4	29
5.6	9.0	10.1	3.4	5.6	30
9.0	4.5	9.0	4.5	4.5	28
7.9	5.6	9.0	2.3	3.4	25
3.4	5.6	6.8	5.6	2.3	21
6.8	0.0	5.6	7.9	4.5	22
6.8	4.5	3.4	2.3	5.6	20
2.3	6.8	2.3	7.9	10.1	26
4.5	6.8	3.4	10.1	9.0	30
6.8	5.6	7.9	10.1	7.9	34
9.0	7.9	6.8	9.0	7.9	36
111	107	113	98	104	533

Figure 7: Heatmap indicating the percentage of participants choosing the word in each cell for group "20"

29.6	32.1	27.2	30.9	40.7	130
28.4	17.3	24.7	17.3	32.1	97
28.4	19.8	25.9	38.3	35.8	120
34.6	34.6	37.0	32.1	33.3	139
98	84	93	96	115	486

A small bias favours of the lowest lines of the array, as well as top line, with little horizontal bias (the choosing bias is around 11% above average for group "20", with $p < 0.02$, and 26% for group "100", with $p < 0.03$, both analyses done with ANOVA).

4.2 Memorization

After choosing a passphrase and performing a distractor task, participants were asked to recall their original passphrase; 46% of participants recalled all 6 words in their passphrase without any errors. An additional 20% remembered their words but made a typo.

The table in Figure 8 sums up the recall success rate and types of errors when recalling passphrases. A full explanation is given below. The rates are calculated separately for group "20", group "100", and the control group. They are also split between the first and the second section, in which (except for the control group) participants were reminded of their original array of words.

The following mistakes and errors showed up in participant recall of newly chosen passphrases:

Figure 8: Total number of errors by type

Section	Correct	Typo	Variant	Order	Miss	Wrong
1:20	19/47	6	8	6	26	5
1:100	26/51	10	5	3	16	4
Control	6/26	11	11	10	31	12
2:20	14/29	1	2	8	0	3
2:100	15/26	4	2	3	1	4

- Correct shows the proportion of participants who wrote the passphrase perfectly[2].
- Typos are simple one-letter errors or exchanges between two adjacent letters.
- Variants are like typos in that they are related English words. Most of those were verbs where the participant added or removed an 's' (or less frequently an 'ed' or 'ing').
- Orders are errors where at least two words are exchanged in the passphrase.
- Misses are words that are entirely missing from the passphrase entered.
- Wrong words are ones that have no relation to any word in the original passphrase.

Overall, 94% of the 51 people that were guided with "100" words remembered at least 5 of their words, and 69% remembered all their words but made a small mistake (like getting them in the wrong order). 81% of the 47 people who were shown 20 words remembered at least 5 of their words, and 64% made at most simple mistakes. The control group demonstrated lower performance, with 38% remembering 5 of their words, and 27% making only simple mistakes.

When comparing the number of people who correctly remembered their whole passphrase, group "100" is superior to the control group ($p < 0.02$). This effect is magnified when comparing not the participants but the words directly. The words in passphrases made from the 100 words array were better remembered than those made from the 20 words array ($p < 0.03$), with only 0.4 words forgotten or wrong per participant against 0.7. Similarly, the words in passphrases made from the 20 words array were themselves much better remembered than the ones given to participants in the control group ($p < 10^{-4}$), who had an average of 1.3 words forgotten.

The creation of sentence-like passphrases had no statistically significant impact on overall success rate in either group, with a small decrease in misses and a increase in false words ($p > 0.05$).

Table 9 shows the error analysis restricted to those who remembered the passphrase correctly just after making it in the first exercise (the one asking them to type the first two letters of each word).

Remarks. The preceding error tables do not take into account four anomalous behaviours. Two participants (one in each group) made a typo in their original passphrase (phrases were only counted as correct when typed without the typo). One participant, when asked for the first two letters for each word in their passphrase, typed random letters on their keyboard, and one typed something that looked like the requested twelve-character string with lots of mistakes. Both of those were in group "20" and were not counted in the analysis. One participant in group "100" also double-clicked on the next button and was taken directly to the second try and shown their array of words. Finally, four participants in the control group showed no attempt to recall their passphrase, responding with random words (and in one case not even filling the 6 phrase positions), and were removed from the dataset. Including these would have only strengthened the results that guided choice helps.

[2]In the table, there is one more participant per group in the second section than there should be. This is due to one participant in group "100" double-clicking the submit button and getting directly to the second section, and one participant in group "20" writing nonsense in the first section but getting five correct words in the second.

Figure 9: Errors by type for participants with correct first exercise

Section	Correct	Typo	Variant	Order	Misses	Wrong
1:20	19/41	4	8	2	14	5
1:100	26/45	9	5	1	14	1
Control	6/15	5	4	1	7	2

Language. People that identified their primary language as English were balanced between the two groups (25 of 51 in group "100" and 26 of 47 in the other). Language did not show a statistically significant effect: 21 out of 51 people who indicated English as their primary language were correct on the first try, as were 23 out of the 48 who indicated another language. Primary English speakers had more misses (29 against 13) and an equal share of wrong words.

Time. No statistically significant advantage was shown for participants who spent more time designing their passphrase. People who recalled their words perfectly appeared to take 5-10% longer on average, but 10-15% less time for the median, showing no clear effect.

Some of the participants disabled the JavaScript functions needed to record the time taken[3].

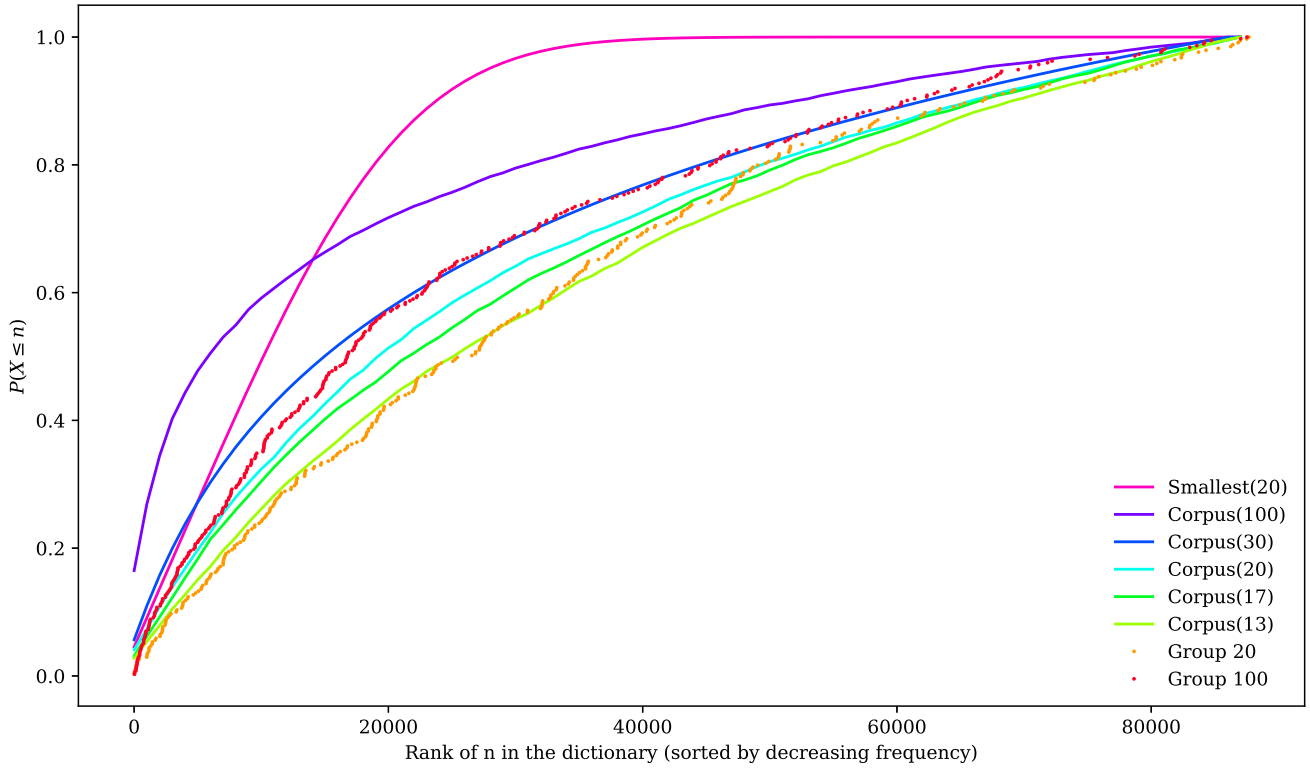
4.3 Guessing

Most participants only tried to guess 1 other passphrase, but 21 participants tried to guess between 2 and 4 passphrases. Words that were in the original passphrase with a minor modification (such as a typo) were counted as correct in this exercise, like those that were only in the wrong place.

On average, participants succeeded at guessing 0.85 of 6 words chosen by another person from 100 word arrays, and 2.15 of 6 words from 20 word arrays. This is significantly higher than a random guess (which would on average get 0.36 and 1.80 words correct), but not better than (an educated guess) focusing on common words or positions. With access to the array, positional guessing would get respectively 0.63 and 2.21 words correct. Purely semantic guessing would get respectively 1.40 and 2.93 words correct.

[3]These are disabled by default on most Apple iPhones.

Figure 10: Cumulative distribution function of indices of words chosen for both groups and 6 models



5 STATISTICAL MODELLING

5.1 Strategies and entropy

The effect of participant choice on the entropy of passphrases was tested. Models simulated in Python were used to analyse the word choice of participants when presented with arrays of words, with three main strategies:

The *Smallest*(n) strategy corresponded to picking the six most frequent words presented in the array of n words.

The *Uniform*(n) strategy being equivalent to sampling random words uniformly from a dictionary of size n, entropy was computed exactly. The table in Figure 11 shows the entropy per word for a few strategies. The *87,691* here corresponds to our described dictionary of *87,691* words, and *300,000* to Norvig's more complete dictionary of *300,000* words.

Finally, the *Corpus*(n) strategy corresponded to picking each word from an array of n words according to a distribution where the probability to pick each word w is a function of $f(w)$, its frequency in the language. This corresponds to models for how the distribution of words is biased in many different corpora of texts. The model used here is Zipf's law[Ha et al. 2002], stating that the probability of choosing a word $p(w)$, is inversely proportional to its rank in the frequency list:

$$p(w) \propto \frac{1}{rank(w)}$$

Informally, the 100^{th} most used word is chosen with a frequency about twice the frequency of the 200^{th} most used word.

10^9 simulations were run for both 20-word and 100-word arrays with each strategy to estimate the probabilities – and the entropy[4]. For entropy E, the formula to compute it is:

$$E = -\sum_i p_i ln(p_i)$$

In this formula, p_i is the position in an ordering of word frequency that the word occupies in the array of words.

Figure 11: Strategies and entropy

Strategy	Entropy (bits)	Strategy	Entropy
Uniform(87,691)	16.42	*Smallest*(20)	12.55
Corpus(13)	16.25	*Uniform*(5,000)	12.29
Corpus(17)	16.15	*Uniform*(2,000)	10.97
Corpus(20)	16.10	*Smallest*(100)	10.69
Corpus(30)	15.92	*Corpus*(300,000)	8.94
Corpus(100)	15.32	*Corpus*(87,691)	8.20
Uniform(10,000)	13.29		

A much bigger sample would be needed to exactly compute the entropies of user behaviours. However, experimental entropy can

[4]The error bounds due to the simulations are quite smaller than 0.01 bits.

be bounded by using distributions for known entropies. As such, the cumulative distribution functions for the experimental groups and models were computed. Although they do not make direct strategic sense, we included *Corpus*(17) and *Corpus*(13) in Figure 10, as they bound the observed curve for group "20".

Experimental values for group "20" are slightly above *Corpus*(17) around the 50,000$^{\text{th}}$ word. An upper bound of *Corpus*(20) could be chosen, but there is a strong argument for using *Corpus*(17). This is more affected by the values of the high p_i, as the function changes less as it gets to the least common words (making it concave). As the p_is shown in Figure 10 are also sorted in decreasing order, it means a small bump in word choice in the first part of the curve is more than compensated by the lack of a bump in the second part. As such, it is reasonable to infer that the entropy corresponding to participants' behaviours in this group is between *Corpus*(13) and *Corpus*(17).

A slightly tighter fit can be obtained by taking not the simplest Zipf's formula but the more general one, with, for $\beta > 1$:

$$p(w) \propto \frac{1}{(rank(w))^{\beta}}$$

In such a case, setting $\beta = 1.35$ makes *Corpus-Zipf*(13) a tighter fit than previous curves, giving an entropy of 16.19 bits[5]. However, the presence of noise in the data means that a search for a more accurate model would be premature.

This analysis shows that six-word passphrases generated with the method proposed and a 100-word array have at least 95 bits of entropy, and ones created with a 20-word array have nearly 97 bits of entropy.

5.2 Semantic aspects

This section has focused on the frequency of words as a proxy for their familiarity. This ignores other possibilities such as emotional attachment to certain words, linked to the particularities of each user. For example, a dog owner might choose the word dog if it appeared in the array. We can show here that the magnitude of such an effect should be quite low.

Let's suppose that each user has an set of 100 words that they will automatically choose whenever they appear. They have the opportunity of choosing such a word with probability at most

$$1 - (1 - \frac{100}{87685})^{100} \approx 11\%$$

An adversary with the word set could, for each position, try this set of words and fill the rest with a dictionary[6], lowering the number of possibilities to test. With probability 0.11, such an adversary could then reduce the total entropy from 95 bits to 89 bits.

Moreover, many emotionally-loaded words such as "dog", "love" or "president" are already among the most frequent words (the ones given are all in the top 1,000). As such, people might choose them with a high probability, no matter their individual preferences, so the marginal information – and the corresponding entropy loss – should be even lower than the bound already given.

[5]One could also use a Zipf-Mandelbrot model [Oldfield 1968], but the additional parameter would be hard to validate accurately without having a sample of at least 10,000 participants

[6]Getting a single word from the set is already a low-probability event, and getting more than one happens with probability at most $2 * 10^{-4}$.

6 LIMITATIONS

6.1 Ecological validity

As this was an online study done in the wild, it did not happen in a controlled laboratory environment, anomalous participant behaviour could not be detected. The main problem with this design might come from users writing down their passwords somewhere, affecting the memorability results. Two arguments help us believe that this had little to no impact on the results. The first comes from the fact that writing down one's passphrase should take a certain amount of time, and there was no demonstrated correlation between the speed during the creation and the ability to remember one's passphrase correctly. The second comes from [Yang et al. 2016], where participants in the study were encouraged to create passphrases and use any technique they generally used to remember passwords and passphrases. Despite knowing in advance that they would have to remember the passphrase for a week, over 80% of participants reported not having written down their passphrase.

6.2 Short-term and long-term memory

As the users were not asked any identifying information, it was not possible to ask them to return to the experiment, to estimate the effects of the method on long-term memorability of passphrases. The main reason we added the distractor task in the experimental protocol was to disrupt participant's short-term memory to look at long-term memory effects. We compared recall rates with the long-term recall rates for the variety of passphrase creation strategies shown in [Yang et al. 2016]. The after-distractor memorability observed here is closer to the long-term memorability they observed. This is consistent with the fact that no variation in recall rates between strategies was observed in their short-term experiment, unlike their long-term experiment and our data. Moreover, as the participants had no intrinsic or financial motivation to recall their passwords, their recall rates should be lower than in real world usage.

6.3 Free choice of words

This study did not include a fourth group who were free to choose words in any way they wanted. This could have been interesting in the goal of comparing memorability but two factors motivated the absence of this second sort of control group. The first is that we could not ensure that people would pick new sentences and not ones they were already trained on, which would skew the results. More importantly, we believe that the behaviours observed in [Yang et al. 2016] indicate that free word choice is too much of a security concern, and not a viable option.

7 DISCUSSION

The above results demonstrate that passphrases created by choosing words from an array of random words are more memorable than automatically generated ones. While past studies have shown that choosing familiar words for passphrases led to huge entropy reductions, our technique obviated this. The entropy cost due to the choice in our system is negligible, staying between 1% and 3% depending on array size. The method also allows the use of a large dictionary to choose known words from, leading to much higher

entropy per word in the end. The long-term memory performance can only be intuited at, but the magnitude of the effect when compared with alternatives on short-term memory with a distractor task should be a good indicator.

Multiple surprising behaviours were observed, confirming certain hypotheses and refuting others. Firstly, the participants' choices were influenced by the positions of the words in the arrays presented to them. In group "20", this led 41% of them to choose the word in the upper left corner, instead of the expected 30%. Variations of a factor 2 between different lines of the array were found in group "100". There was a significant bias in favour of the last lines, and a smaller one for the top line, with no significant horizontal effect[7].

The tendency to choose familiar words was stronger than the positional bias, although the linguistic bias was still weaker than in the English language (as predicted by Zipf's law). We observed that *Corpus*(13-17) might be better fits than *Corpus*(20) for group "20". This can be explained by the fact that its participants took the second line of the array at a lower frequency. Similarly, positional effects could partially explain why the distribution is closer to *Corpus*(30) than *Corpus*(100).

One might expect participants to use mnemonics and create sentence-like passphrases that used common patterns, but the only syntactic pattern that appeared more than thrice was 8% of participants choosing three nouns in a row. As nouns form the bigger share of the dictionary used, even those passphrases are secure, reducing the entropy by at most 5 bits (out of more than 96 bits of entropy). Having more words to choose from did not increase the tendency to create syntactically correct sentences but reduced it instead.

The position bias should not be seen as a weakness of the system but as another source of its security: as the position is truly random, a stronger positional bias leads to a more uniform – and higher-entropy – passphrase. This is true as long as the adversary cannot get the initial word array, but in such a case the system is already unsecure.

Word choice patterns held across the range of proficiencies in English. Those results are true not just for primary English speakers but also for people for whom it is a second language, with small differences in word choice and no significant difference in memorability.

The task where participants were supposed to guess each other's words had a purpose beyond distracting and impairing their memory. We were hoping to find whether a simple strategy could explain the participants' choices, in which case some would get very good results. The absence of such successful participants shows that if a general strategy to explain participants' choices exists, it has eluded both us and them.

8 CONCLUSION

This paper shows that guiding people to choose from an array of random words can build a high-entropy, more memorable passphrase. Choosing from an array successfully stops people from choosing easy-to-guess favourite words. The reduction in entropy when compared with random passphrase generation is offset by the improved usability and memorability that come with choice. Since people can choose words they know, larger dictionaries can be used. Final entropy by word then exceeds the levels of random passphrase generation by 20 – 30%, reaching 95 and 98 bits of entropy.

The main, but weak, predictor of word choice from the random word array was how familiar a word was to the participant, for both primary and non-primary speakers. However, this bias was weaker than expected and letting people choose from an array gave between 97% and 99% of the maximal entropy achievable, depending on the size of the array.

Bigger array size (giving more choice to the participant) was linked to improved memorability of the passphrase. With a 100-word array, 94% remembered at least 5 of their 6 words the first time they were asked to recall, despite performing a distractor task. With actual use, remembering the passphrase should be easy.

The advantage of selecting from an array of random words is that it gives users an easy way to create secure and memorable passphrases, with only a random generator and a dictionary. Security of the guiding array is achieved by only generating it when needed, locally on the user's machine. Generating a list which would then be sent over the network is highly inadvisable as it drastically lowers the entropy, even more so if the adversary knows the position of the words. However, this needn't be a problem in practice as the secret array to select from can be produced on any machine, the dictionary itself and the code needed taking less than 300KB of memory. Although it could be of help if the user forgets their passphrase, storage of the array itself on a local machine once used would introduce real vulnerabilities. This method should only be used at a user's initiative – being a tool to help them make better passphrases – to avoid the same pitfalls as counter-productive password constraints.

This work demonstrates that large improvements can be achieved in passphrase usability while increasing their entropy. Below are a few suggestions for usability/security questions left to test for the guided passphrase scenario:

- How well does the short-term memorability with a distractor task predict long-term memorability for passphrases?
- Can putting high-frequency words closer to the middle of the array compensate for the positional advantage?
- Can other visual presentations, such as word clouds, make word choice from the presented set even more uniform?
- Would the high uniformity of word frequency be as or even more successful with even larger dictionaries, for example with the full SOWPODS and its 276663 words?
- Why is choosing from 100 words more memorable than choosing from 20? It isn't because people took longer, as they didn't. It might be as it gives users more personalized choices or the ability to select more familiar words. It might also be for some not yet explored reason. Does memorability continue to increase with arrays of more than 100 words? What is the nature of the trade-off, and must there be a compromise between entropy and memorability?
- How does the size of the array affect reading patterns and word choice? Does the left-wise bias in group "20" come from different reading patterns?

[7] We did not compare multiple ways of presenting a word field. While a linear presentation of the choice words was not tested, it would be likely be much less user-friendly.

- A significant portion of the errors, both in this study and previous ones [Yang et al. 2016], come from simple errors like typos or an added 's'. Could one make an error-tolerant passphrase authentication system that would be compatible with the usual hash-and-compare techniques?
- Are there even better ways to make highly-memorable high-entropy passphrases?

Making easy-to-remember passphrases, such as presented here, should improve security in many use-cases. With these demonstrations, we hope to inspire more work that will make secure passphrases that are as easy to remember as the song you can't get out of your head.

REFERENCES

Alan Baddeley and Graham James Hitch. 1974. *Working memory*. Vol. 8. Academic Press, 47–90.

Nicolas Blanchard, Leila Gabasova, Ted Selker, and Eli Sennesh. 2018. Cue-Pin-Select, a Secure and Usable Offline Password Scheme. (July 2018). https://hal.archives-ouvertes.fr/hal-01781231 working paper or preprint.

Jeremiah Blocki, Manuel Blum, Anupam Datta, and Santosh Vempala. 2014. Towards Human Computable Passwords. *arXiv preprint arXiv:1404.0024* (2014).

Niels J Blunch. 1984. Position bias in multiple-choice questions. *Journal of Marketing Research* (1984), 216–220.

J. Bonneau. 2012. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *2012 IEEE Symposium on Security and Privacy*. 538–552. https://doi.org/10.1109/SP.2012.49

Joseph Bonneau and Stuart E Schechter. [n. d.]. Towards Reliable Storage of 56-bit Secrets in Human Memory.

Joseph Bonneau and Ekaterina Shutova. 2012. Linguistic properties of multi-word passphrases. In *International Conference on Financial Cryptography and Data Security*. Springer, 1–12.

Sacha Brostoff and M. Angela Sasse. 2000. *Are Passfaces More Usable Than Passwords? A Field Trial Investigation*. Springer London, London, 405–424. https://doi.org/10.1007/978-1-4471-0515-2_27

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology* 7 (2016), 1116. https://doi.org/10.3389/fpsyg.2016.01116

Kai Cao and Anil K. Jain. 2016. Hacking Mobile Phones Using 2D Printed Fingerprints.

Lorrie Faith Cranor. 2014. What's wrong with your pa$$word. https://www.ted.com/talks/lorrie_faith_cranor_what_s_wrong_with_your_pa_w0rd.

Lorrie Faith Cranor. 2016. Time to rethink mandatory password changes. https://www.ftc.gov/news-events/blogs/techftc/2016/03/time-rethink-mandatory-password-changes

Leilei Gao and Itamar Simonson. 2016. The positive effect of assortment size on purchase likelihood: The moderating influence of decision order. *Journal of Consumer Psychology* 26, 4 (2016), 542–549.

S. Garfinkel and H.R. Lipford. 2014. *Usable Security: History, Themes, and Challenges*. Morgan & Claypool Publishers. https://books.google.fr/books?id=HPS9BAAAQBAJ

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F. J. Smith. 2002. Extension of Zipf's Law to Words and Phrases. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1 (COLING '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–6. https://doi.org/10.3115/1072228.1072345

George W Hartmann. 1946. Further evidence on the unexpected large size of recognition vocabularies among college students. *Journal of educational psychology* 37, 7 (1946), 436.

Yasser M. Hausawi and William H. Allen. 2014. *An Assessment Framework for Usable-Security Based on Decision Science*. Springer International Publishing, Cham, 33–44. https://doi.org/10.1007/978-3-319-07620-1_4

Charles Hulme, Steven Roodenrys, Richard Schweickert, Gordon DA Brown, Sarah Martin, and George Stuart. 1997. Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23, 5 (1997), 1217.

Blake Ives, Kenneth R. Walsh, and Helmut Schneider. 2004. The Domino Effect of Password Reuse. *Commun. ACM* 47, 4 (April 2004), 75–78. https://doi.org/10.1145/975817.975820

Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Serge Egelman. 2011. Of Passwords and People: Measuring the Effect of Password-composition Policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2595–2604. https://doi.org/10.1145/1978942.1979321

JH Krantz. 1998. Psychological research on the net. *WWW document* (1998).

Cynthia Kuo, Sasha Romanosky, and Lorrie Faith Cranor. 2006. Human selection of mnemonic phrase-based passwords. In *Proceedings of the second symposium on Usable privacy and security*. ACM, 67–78.

Micah Lee. 2015. Passphrases that you can memorize – but that even the NSA can't guess. https://theintercept.com/2015/03/26/passphrases-can-memorize-attackers-cant-guess.

Kristina Lerman and Tad Hogg. 2014. Leveraging position bias to improve peer recommendation. *PloS one* 9, 6 (2014), e98914.

Peter Lipa. 2016. The Security Risks of Using "Forgot My Password" to Manage Passwords. https://www.stickypassword.com/blog/the-security-risks-of-using-forgot-my-password-to-manage-passwords. Accessed: 2017-12-18.

Wanli Ma, John Campbell, Dat Tran, and Dale Kleeman. 2010. Password entropy and password quality. In *Network and System Security (NSS), 2010 4th International Conference on*. IEEE, 583–587.

Jim Marquardson. 2012. Password Policy Effects on Entropy and Recall: Research in Progress. (2012).

George A Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

Peter Norvig. 2009. Natural language corpus data. *Beautiful Data* (2009), 219–242.

Richard Charles Oldfield. 1968. *Language: selected readings*. Vol. 10. Penguin.

L Payne Stanley. 1951. The art of asking questions.

Denise Ranghetti Pilar, Antonio Jaeger, Carlos F. A. Gomes, and Lilian Milnitsky Stein. 2012. Passwords Usage and Human Memory Limitations: A Survey across Age and Educational Background. *PLoS One* 7, 12 (05 Dec 2012), e51067. https://doi.org/10.1371/journal.pone.0051067 PONE-D-12-21406[PII].

Sigmund N Porter. 1982. A password extension for improved human factors. *Computers & Security* 1, 1 (1982), 54–56.

Stefan Rass and Sandra König. 2018. Password Security as a Game of Entropies. *Entropy* 20, 5 (2018), 312.

P. Venkata Reddy, Ajay Kumar, S. Rahman, and Tanvir Singh Mundra. 2008. A New Antispoofing Approach for Biometric Devices. *IEEE transactions on biomedical circuits and systems* 2 4 (2008), 328–37.

Virginia Ruiz-Albacete, Pedro Tome-Gonzalez, Fernando Alonso-Fernandez, Javier Galbally, Julian Fierrez, and Javier Ortega-Garcia. 2008. *Direct Attacks Using Fake Images in Iris Verification*. Springer Berlin Heidelberg, Berlin, Heidelberg, 181–190. https://doi.org/10.1007/978-3-540-89991-4_19

Sean M. Segreti, William Melicher, Saranga Komanduri, Darya Melicher, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. 2017. Diversify to Survive: Making Passwords Stronger with Adaptive Policies. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, Santa Clara, CA, 1–12. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/segreti

Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip (Seyoung) Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2014. Can Long Passwords Be Secure and Usable?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2927–2936. https://doi.org/10.1145/2556288.2557377

Koen Simoens, Pim Tuyls, and Bart Preneel. 2009. Privacy weaknesses in biometric sketches. In *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 188–203.

Daniel F. Smith, Arnold Wiliem, and Brian C. Lovell. 2015. Face Recognition on Consumer Devices: Reflections on Replay Attacks. *IEEE Transactions on Information Forensics and Security* 10 (2015), 736–745.

Mariam M Taha, Taqwa A Alhaj, Ala E Moktar, Azza H Salim, and Settana M Abdullah. 2013. On password strength measurements: Password entropy and password quality. In *Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on*. IEEE, 497–501.

Umut Topkara, Mikhail J. Atallah, and Mercan Topkara. 2007. Passwords Decay, Words Endure: Secure and Re-usable Multiple Password Mnemonics. In *Proceedings of the 2007 ACM Symposium on Applied Computing (SAC '07)*. ACM, New York, NY, USA, 292–299. https://doi.org/10.1145/1244002.1244072

Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. 2016. Understanding Password Choices: How Frequently Entered Passwords Are Re-used across Websites. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 175–188. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/wash

Jeff Yan, Alan Blackwell, Ross Anderson, and Alasdair Grant. 2004. Password Memorability and Security: Empirical Results. *IEEE Security and Privacy* 2, 5 (Sept. 2004), 25–31. https://doi.org/10.1109/MSP.2004.81

Weining Yang, Ninghui Li, Omar Chowdhury, Aiping Xiong, and Robert W. Proctor. 2016. An Empirical Study of Mnemonic Sentence-based Password Generation Strategies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. ACM, New York, NY, USA, 1216–1229. https://doi.org/10.1145/2976749.2978346

Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th international conference on World wide web*. ACM, 1011–1018.